Debiasing NLU Models via Causal Intervention and Counterfactual Reasoning

Bing Tian[1], **Yixin Cao**[2], **Yong Zhang**[1*], **Chunxiao Xing**[1]

[1]DCST, BNRist, RIIT, Institute of Internet Industry, Tsinghua University, Beijing, China
[2]Singapore Management University
tb17@mails.tsinghua.edu.cn, yixin.cao@ntu.edu.sg, {zhangyong05, xingcx}@tsinghua.edu.cn

Abstract

Recent studies have shown that strong Natural Language Understanding (NLU) models are prone to relying on annotation biases of the datasets as a shortcut, which goes against the underlying mechanisms of the task of interest. To reduce such biases, several recent works introduce debiasing methods to regularize the training process of targeted NLU models. In this paper, we provide a new perspective with causal inference to find out the bias. On the one hand, we show that there is an unobserved confounder for the natural language utterances and their respective classes, leading to spurious correlations from training data. To remove such confounder, the backdoor adjustment with causal intervention is utilized to find the true causal effect, which makes the training process fundamentally different from the traditional likelihood estimation. On the other hand, in inference process, we formulate the bias as the direct causal effect and remove it by pursuing the indirect causal effect with counterfactual reasoning. We conduct experiments on large-scale natural language inference and fact verification benchmarks, evaluating on bias sensitive datasets that are specifically designed to assess the robustness of models against known biases in the training data. Experimental results show that our proposed debiasing framework outperforms previous state-of-the-art debiasing methods while maintaining the original in-distribution performance.

Introduction

Despite the impressive performance on many NLU benchmarks (Wang et al. 2019), recent studies have demonstrated that neural models tend to rely heavily on existing annotation biases, without learning the underlying task (Gururangan et al. 2018; Poliak et al. 2018; Schuster et al. 2019; McCoy, Pavlick, and Linzen 2019; Shah, Schwartz, and Hovy 2020; Zhang et al. 2021). These biases are commonly characterized as surface features of input examples that are strongly associated with the target labels in the datasets. For instance, natural language inference (NLI) is a task to conduct combined feature reasoning to determine whether a hypothesis sentence can be inferred from a premise sentence (Dagan, Glickman, and Magnini 2005). However, recent work has demonstrated that large-scale NLI benchmarks contain annotation artifacts: *e.g.*, the entailed hypotheses tend to replace exact numbers/gender with approximates/generic words (some, at least, human, people etc.), purpose clauses are a sign of neutral hypotheses, and negation is correlated with contradiction label (Gururangan et al. 2018). Table 1 shows the examples from the widely used SNLI dataset that demonstrate these phenomena. With only processing the hypothesis, models can reach accuracy scores as high as twice the majority baseline (67% vs.34%) when predict the class within the SNLI dataset (Tsuchiya 2018; Gururangan et al. 2018). Fact verification (FEVER), another NLU task suffers from similar issue: a claim-only BERT (Devlin et al. 2019) model that classifies each claim on its own, without associated evidence achieves 61.7%, far above the majority baseline (33.3%) (Schuster et al. 2019). As a result, models will fail to generalize well if they simply memorize the statistical shortcuts during training, and suffer from a huge drop in performance when evaluated on the datasets that are carefully designed to limit the spurious cues.

Premise and Hypothesis	Label
P: A woman is talking to two men. H: There are **at least** three **people**.	entailment
P: Two dogs are running through a field. H: Dogs are running **to catch a stick**.	neutral
P: The woman is awake. H: The woman is **not** awake.	contradiction

Table 1: Examples from SNLI that illustrate the annotation artifacts.

Recent popular solution to such issue is to develop debiasing mehods that overcome these biases at the training stage (Belinkov et al. 2019; He, Zha, and Wang 2019; Stacey et al. 2020; Mahabadi, Belinkov, and Henderson 2020; Utama, Moosavi, and Gurevych 2020; Ghaddar et al. 2021). Namely, they first use a bias model [1] to identify biased samples. And then adversarial learning or ensemble training are utilized to either remove the bias from sentence encoder or control the training loss by discouraging learning from the bias samples. However, we argue that biased

*corresponding author

[1]We follow the terminology used by (He, Zha, and Wang 2019)

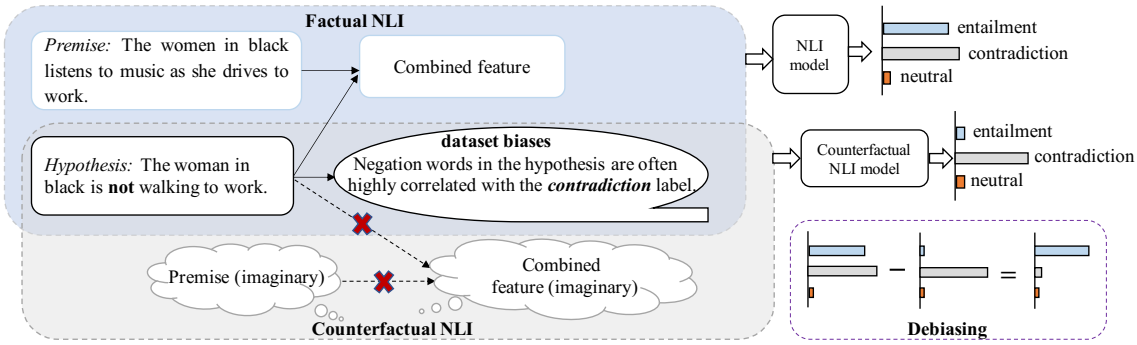


Figure 1: An illustration of factual and counterfactual NLI, as well as the debiasing strategy. Factual NLI depicts the fact where model sees the hypothesis and extracts the combined feature of premise and hypothesis. Counterfactual NLI means that model sees the hypothesis but the combined feature and premise are coming from the imagined world.

samples can include some information necessary to perform the NLU tasks and learning from these examples helps the model maintain the original in-distribution accuracy. Unlike previous works, in this paper, we hand over the debiasing to the inference stage, and focus on how to disentangle the learned general features and memorized dataset bias.

Specifically, motivated by causal inference (Didelez and Pigeot 2001; Pearl 2013), we propose a novel bias mitigation strategy from a causal-effect look. We first formulate the procedure of NLU tasks from the causal view with a Structural Causal Model (SCM) (Pearl et al. 2000). In the training process, we believe that there exists an unobserved confounder representing the *annotation preference* between natural language utterances and their respective labels, leading to the spurious correlation. Recall the process of the corpus generation of NLI, crowd workers are presented with a premise p and are required to generate three new sentences (hypotheses) that entails, contradicts, or is logically neutral with respect to the premise. This protocol makes the preference of annotators inevitably affect both the hypothesis and the label (Gururangan et al. 2018). Figure 2 illustrates how such annotation preference confounds the generation process of hypothesis sentences. On the Structural Causal Model (SCM), the confounder appears in the undesired causal path ($a.k.a.$, backdoor (Pearl et al. 2009)), which is an indirect causal link from input H to output L: $H \leftarrow U \rightarrow L$. Considering such issue, we apply deconfounded training with causal intervention and use the do-calculus $P(L|do(H))$ (Pearl, Glymour, and Jewell 2016) to calculate the causal effect, which is fundamentally different from the conventional likelihood $P(L|H)$.

Then in the inference stage, we further formulate the hypothesis/claim only bias as the direct causal effect of hypothesis/claim on labels, and conduct the debiasing by subtracting the direct causal effect from the total causal effect. To reach this goal, one question needs to be answered: How could we estimate the causal effects in NLU tasks?

For this question, inspired by recent counterfactual reasoning works (Qian et al. 2021; Niu et al. 2020), we introduce two situations of inference process: factual NLI and

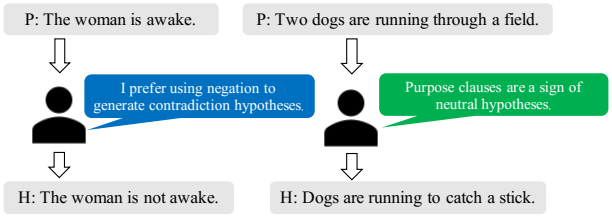


Figure 2: Illustration of the annotation preference.

counterfactual NLI, to obtain the total causal effects and the direct causal effect respectively. Figure 1 [2] illustrates these two situations. Intuitively, factual NLI depicts the situation where both P and H are available. In this case, we can estimate the total causal effect of P and H on L. However, factual NLI cannot disentangle the annotation bias and the needed combined feature reasoning. Therefore, we consider the following counterfactual question: "What will the prediction be if seeing the hypothesis sentence only and had not seen the premise and the combined feature?" Under this imaginary situation, since the effects of P and mediator C are both blocked, NLI models can only make decisions rely on the hypothesis-only impact. Therefore, the bias can be identified by estimating the direct causal effect of H on L under the situation of counterfactual NLI.

We conduct experiments on large-scale NLI and fact verification benchmarks, evaluating on bias sensitive datasets SNLI-hard and fact verification symmetric test set (Gururangan et al. 2018; Schuster et al. 2019) that are specifically designed to assess the robustness of models against known biases in the training data. Experimental results show that our Causal Intervention and Counterfactual Reasoning (CICR) based framework outperforms existing debiasing methods by large margins on the bias sensitive datasets, and remains stable on the original SNLI and FEVER datasets.

Contributions of this paper are summarized as following:

[2] Here, we take natural language inference as an example. Other NLU tasks such as fact verification can be analogous in a similar way.

- We formulate biases in tasks of NLI and fact verification from the new causal view with a SCM.
- Based on the causal framework, we conduct causal interventions via backdoor adjustment to remove spurious correlations introduced by the annotation preference confounder. For mitigating the bias, we propose a counterfactual reasoning framework to pursue the indirect causal effect as the inference logits.
- We conduct extensive experiments on large-scale NLI and fact verification benchmarks, and achieve state-of-the-art debiasing results while maintaining the original in-distribution performance.

Related Work

Debiasing NLU Models

Recent solutions to reduce the dataset biases in NLU tasks can be grouped into three categories: data augmentation (Gururangan et al. 2018; Schuster et al. 2019; Kaushik, Hovy, and Lipton 2020; Nie et al. 2020), adversarial training (Belinkov et al. 2019; Stacey et al. 2020) and model ensembles (Clark, Yatskar, and Zettlemoyer 2019; He, Zha, and Wang 2019; Mahabadi, Belinkov, and Henderson 2020; Utama, Moosavi, and Gurevych 2020; Ghaddar et al. 2021). First, Nie et al. (2020) proposed an iterative human-and-model-in-the-loop solution for NLU dataset collection targeting at addressing robustness issues in existing datasets. Kaushik et al. (2020) designed a new dataset creation procedure in which humans counterfactually provided labels and intervened upon the data. To allow proper evaluation, recent studies have tried to create new evaluation datasets that do not contain idiosyncratic biases (Gururangan et al. 2018; Schuster et al. 2019; McCoy, Pavlick, and Linzen 2019). Second, adversarial learning is utilized to remove the hypothesis-only bias from models trained on SNLI. Specifically, Belinkov et al. (2019) and Stacey et al. (2020) discouraged the hypothesis encoder from learning the biases by designing a classifier trained to learn the bias from sentence representations, which in turn update to reduce the performance of the bias classifier in an adversarial manner. Third, ensemble-based methods (Clark, Yatskar, and Zettlemoyer 2019; He, Zha, and Wang 2019; Mahabadi, Belinkov, and Henderson 2020; Utama, Moosavi, and Gurevych 2020; Ghaddar et al. 2021) proposed to use a separated biased model, which is a weak classifier that is trained using only features that are known to be insufficient to perform the task but work well due to biases. The output of this pre-trained biased model is then used to adjust the loss function such that it down-weights the importance of examples that the biased model can solve.

Our work falls into the model ensemble category and the most related work to ours in terms of approach is (Mahabadi, Belinkov, and Henderson 2020). The authors proposed a learning strategy to overcome dataset biases in NLU tasks. They firstly utilized a bias-only branch to leverage biases and shortcuts in the datasets during training. Debiasing strategies then worked by adjusting the cross-entropy loss based on the performance of these bias-only models to down-weight the importance of the biased examples. At the end of the training, they simply removed the bias-only classifier and used the predictions of the base model. In contrast, our learning strategy is designed based on the causal inference. On the one hand, we utilized backdoor adjustments to remove the spurious correlation caused by the confounder. On the other hand, the debiasing is carried out in the inference stage, which can make unbiased decisions with biased observations by removing the direct bias effect.

Causal Inference

Causal inference is the process of determining the independent, actual effect of a particular phenomenon (Pearl et al. 2009), which has been explored for years in psychology, politics and epidemiology (Richiardi, Bellocco, and Zugna 2013; Keele 2015). By removing confounding bias in data, causal inference can provide more reliable explanations and also provide debiasing solutions by learning causal effect rather than correlation effect. Recently, some works (Singh and Sun 2019; Mahajan, Tan, and Sharma 2019; Bengio et al. 2020; Wang et al. 2020) introduced causal inference into machine learning with counterfactual reasoning, trying to endow models the abilities of pursuing the cause-effect. Especially, it has inspired several studies in Natural Language Processing (NLP), including language models (Huang et al. 2020), named entity recognition (Zhang et al. 2021), text classification (Choi et al. 2020; Qian et al. 2021), reading comprehension (Ye, Nair, and Durrett 2021) and vision-language tasks (Teney, Abbasnejad, and van den Hengel 2020; Niu et al. 2020). More recently, a survey introduced the details about causal inference in natural language processing (Feder et al. 2021). Besides, some works focused on data augmentation via generating counterfactual samples to alleviate the spurious correlation issue in sentiment analysis and NLI tasks (Huang, Liu, and Bowman 2020; Yang et al. 2021; Wang and Culotta 2021). Differently, our causal-effect look focuses on counterfactual inference with even biased training data without extra data augmentation. To the best of our knowledge, we are the first to formulate dataset biases in NLI and fact verification as causal effects and debias it based on counterfactual reasoning framework.

Preliminaries

Causal Graph

Causal graph (Pearl, Glymour, and Jewell 2016) is a highly general roadmap specifying the causal dependencies among variables. As shown in Figure 3, it describes how variables interact with each other, expressed by a directed acyclic Bayesian graphical model, $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$ consisting of nodes $\mathcal{N}$ and directed edges $\mathcal{E}$ (*i.e.*, arrows). $\mathcal{N}$ denotes the set of variables, and $\mathcal{E}$ (arrows) represent the causality between two nodes, *i.e.*, $X \rightarrow Y$ denotes that X is the cause and Y is the effect, meaning the outcome of Y is caused by X.

Counterfactual

Counterfactual means "counter to the facts" (Roese 1997), which assigns the "clash of worlds" combination of values to variables. It provides the framework for many statistical procedures intended to estimate causal effects. Take Figure 3(c)

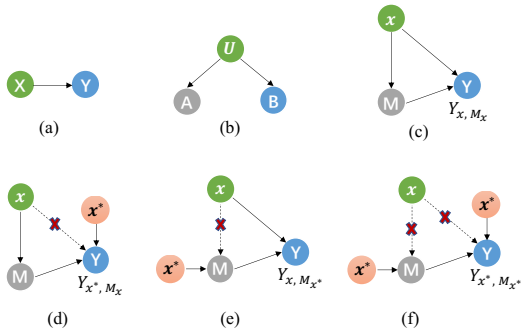


Figure 3: An example of causal graph and counterfactual notations. The dotted line indicates cutting the original incoming arrow.

as an example, in the factual scenario, we have $m = M_x = M(X = x)$. In the counterfactual scenario (Figure 3(d) and (e)), X is set as different values for M and Y. For example, Y_{x^*,M_x} in Figure 3(d) describes the situation where X is set to x^* and M is set to the value when X had still been x, *i.e.*, $Y_{x^*,M_x} = Y(X = x^*, m = M(X = x))$.

Causal Effects

Causal effects are the comparisons between two potential outcomes of the same individual given two different treatments (Rubin 1978). Take Figure 3(c) and (f) as an example, supposed that $X = x$ means we see the X and $X = x^*$ means X is not available. The total effect (TE) of treatment $X = x$ on Y compares these two situations $X = x$ and $X = x^*$, which is denoted as:

$$TE = Y_{x,M_x} - Y_{x^*,M_{x^*}} \tag{1}$$

Total effect can be regarded as the sum of natural direct effect (NDE) and total indirect effect (TIE). NDE represents the effect of X on Y when the mediator M is blocked. It expresses the increase in the outcome Y with X changing from x^* to x under the pure environment M_{x^*}:

$$NDE = Y_{x,M_{x^*}} - Y_{x^*,M_{x^*}} \tag{2}$$

TIE is the difference between TE and NDE, denoted as:

$$TIE = TE - NDE = Y_{x,M_x} - Y_{x,M_{x^*}} \tag{3}$$

In this paper, we use TIE as our unbiased inference results.

Causal Intervention

Causal intervention is utilized to seek the true causal effect of one variable on another when there exists confounders. Figure 3(b) is an example, in which the variable U is the confounder for A and B. In this graph, A does not have causal effect on B because if we only change A and keep U, B will not change. In this case, the confounder makes us cannot use $P(B|A)$ to represent the causal effect, since $P(B|A) - P(B)$ is not always zero. Thanks for the book (Pearl, Glymour, and Jewell 2016), the backdoor adjustments with do-calculus can be used for causal intervention. Specifically, do is a type of intervention, which means that we assign a value to the variable instead of that its parent nodes cause it. When we do a variable, we cut off all the arrows ending to the variable, so that its parents do not cause it any more. For example, in Figure 3(b) , $do(A)$ is that we set variable A as value a while ignoring its caused function (*i.e.*, arrow $A \leftarrow U$). In this way, no confounder will simultaneously cause A and B when calculating $P(B|do(A))$, which means we have de-confounded U.

Methodology

In this paper, we focus on the two tasks of NLI and fact verification, which are considered to contain similar bias namely hypothesis/claim-only bias. In the following introduction, we take NLI as an example, and the fact verification can be analogous in a similar way. Following the common formulation, we regard the NLI task as a multi-class classification problem. NLI models are trained to predict an relationship label from the candidate set $L = l$ given a premise sentence $P = p$ and a hypothesis sentence $H = h$.

De-confounded Training with Causal Intervention

We define the causal graph of NLI in Figure 4(a). Nodes P, H and C represent the input premise, hypothesis and the combined feature of P and H respectively. The final predictive logits L takes inputs from the three branches: the direct effect of the input P and H on L via $P \rightarrow L$ and $H \rightarrow L$, as well as the indirect effect of the input P and H on L via the combined feature C, *i.e.* $C \rightarrow L$.

We donate the score a label l (*i.e.* "entailment") would obtain when the P is set to p (*i.e.* "The women in black listens to music as she drives to work."), H is set to h (*i.e.* "The woman in black is not walking to work.") and C is set to c as:

$$S_{h,p,c}(l)^3 = S(H = h, P = p, C = c), \tag{4}$$

where $c = C_{h,p} = C(H = h, P = p)$. Then the total effect (TE) of the input on label l can be written as:

$$TE = S_{h,p,c} - S_{h^*,p^*,c^*}, \tag{5}$$

where h^* and p^* represent the no-treatment condition where h and p are not given, and $c^* = C_{p^*,h^*}$.

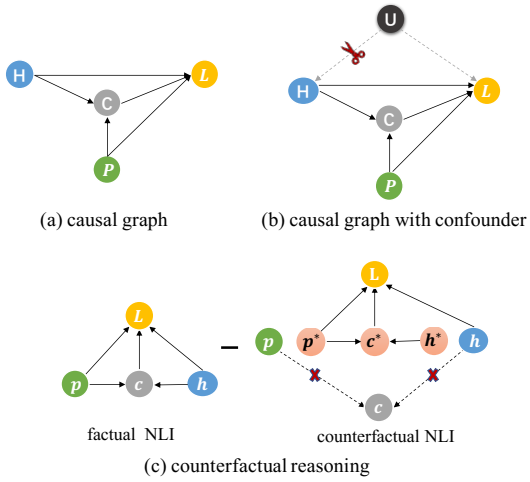
Parameterization Each branch of Figure 4(a) can be formulated as a neural model. And the score $S_{h,p,c}$ is calculated through model ensemble with a fusion function.

$$S_{h,p,c} = \mathcal{F}(S_h, S_p, S_c) \tag{6}$$

where $S_h = \mathcal{F}_H(h)$ is the hypothesis-only branch (*i.e.* $H \rightarrow L$), $S_p = \mathcal{F}_P(p)$ is the premise-only branch (*i.e.* $P \rightarrow L$) and $S_c = \mathcal{F}_{HP}(h, p)$ is the combined feature branch (*i.e.* $C \rightarrow L$). $\mathcal{F}$ is the fusion function to obtain the final score.

In the counterfactual scenery, since the neural models cannot deal with void input, we define the outcome of void input as the same constant a which is a learnable parameter for all the logits. The insight of this setting is that as for human, if we have no information for NLI task, we would like to make

[3]We omit l for simplicity later without loss of generality.



(a) causal graph (b) causal graph with confounder

(c) counterfactual reasoning

Figure 4: Causal graph of NLI and the comparison between factual NLI and counterfactual NLI.

inference by random guess, which means that each candidate label has the same chance to be picked.

We propose two fusion variants, FC (Eq. 7) and SUM (Eq. 8) to combine S_h, S_p and S_c:

$$\mathcal{F}(S_h, S_p, S_c) = \mathbf{W}_c S_c + \mathbf{W}_p S_p + S_h \tag{7}$$

$$\begin{cases} \mathcal{F}(S_h, S_p, S_c) = log\sigma(S_{SUM}) \\ S_{SUM} = S_h + S_p + S_c \end{cases} \tag{8}$$

De-confounder Process with do-calculus As we introduced in previous sections, there exists an unobserved confounder U (Figure 4(b)), which is the cause of spurious correlation between the hypotheses and labels. We believe that such unobserved U should be annotation preference as the process of dataset construction essentially brings humans in the loop. To accomplish the de-confounded training in $H \rightarrow L$ branch, we exploit the backdoor adjustments (Pearl, Glymour, and Jewell 2016) with do-calculus to calculate the corresponding intervention distribution:

$$\begin{aligned} S_h &= P(L|do(H)) \\ &= \sum_u P(L|H,u)P(u|H) \\ &= \sum_u P(L|H,u)P(u) \\ &= \mathbb{E}_u[P(L|H,u)] \end{aligned} \tag{9}$$

Since U is unobserved, we propose learning to approximate it by designing a dictionary D_u as $N \times d$. N is manually set and d is the hidden feature dimension. Here, we set N as 3: the number of labels, so that D_u can be modeled as the annotators' preference over each labels. Note that the NLI is a multi label classification problem, so the last layer of this network is a $softmax$ layer which implements $P(L|H,u)$ as:

$$P(L|H,u) = softmax(g(h,u)) \tag{10}$$

where $g(\cdot)$ is the embedding layer before the $softmax$.

Since Eq. 9 needs expensive samplings for u, we use NWGM approximation (Srivastava et al. 2014; Xu et al. 2015) to efficiently absorb the expectation into the softmax.

$$\begin{aligned} P(L|do(H)) &= \mathbb{E}_u[P(L|H,u)] \\ &= \mathbb{E}_u[softmax(g(h,u))] \\ &\approx softmax(\mathbb{E}_u[g(h,u)]) \end{aligned} \tag{11}$$

In this paper, we model $g(h,u) = W(f(h)+u)$. f is the encoder network to get the embeddings of hypotheses. Then according to the linear additive property of expectation calculation, $\mathbb{E}_u[g(h,u)]$ can be calculated as $W(f(h)+\mathbb{E}_u[D_u])$. In practice, we use a dot-product attention to compute $\mathbb{E}_u[D_u]$. Specifically, $\mathbb{E}_u[D_u] = softmax(L^T K) \odot D_u$, where $L = W_1 h, K = W_2 D_u$ and $\odot$ is the element-wise product, h is the embedding of hypothesis H, and W_1 and W_2 are mapping matrices.

Unbiased Inference with Counterfactual Reasoning

NLI models suffer from the annotation artifacts between the hypotheses and labels, and thus fail to conduct effective combined feature inference. Therefore, we expect NLI models to exclude the direct impact of hypotheses. To achieve this goal, we propose counterfactual NLI to estimate the causal effect of $H = h$ on $L = l$ by blocking the impact of C and P. Counterfactual NLI describes the scenario where H is set to h and C would attain the value c^* when H had been h^* and P had been p^*. Since the response of P and mediator C to inputs is blocked, the model can only rely on the given hypotheses for decision making. Figure 4(c) shows the comparison between factual NLI and counterfactual NLI. We obtain the natural direct effect (NDE) of H on L by comparing counterfactual NLI to the no-treatment conditions:

$$NDE = S_{h,p^*,c^*} - S_{h^*,p^*,c^*} \tag{12}$$

Since the effects of P and C on the labels are blocked, NDE explicitly captures the hypothesis-only bias. Furthermore, the reduction of bias can be realized by subtracting NDE from TE, which is represented as:

$$TIE = TE - NDE = S_{h,p,c} - S_{h,p^*,c^*} \tag{13}$$

We select the answer with the maximum TIE for inference, which is totally different from traditional strategies that is based on the posterior probability *i.e.*, $P(l|h,p)$.

Training and Inference

We follow the training strategy used by (Mahabadi, Belinkov, and Henderson 2020). Figure 5 is an illustration of the training process. Specifically, we jointly optimize the parameters of the base NLI model, the hypothesis-only branch and the premise-only branch using the gradients computed from three losses.

$$Loss_{CE} = Loss_{NLI} + \lambda_H Loss_H + \lambda_P Loss_P, \tag{14}$$

where the main loss $Loss_{NLI}$ refers to the cross-entropy loss associated with the predictions of $\mathcal{F}(S_h, S_p, S_c)$ from Equation 6. We back propagate this loss to optimize all

the parameters Θ_{HPC} which contribute to this loss. Θ_{HPC} is the union of the parameters of the three branches (*i.e.* $H \rightarrow L$, $P \rightarrow L$, and $C \rightarrow L$). In our setup, we share the parameters of the hypothesis and premise encoder among the NLI model, the hypothesis and premise-only branches. The hypothesis and premise-only loss $Loss_H$ and $Loss_P$ are cross-entropy losses associated with the predictions of S_h and S_p from Equation 6. λ_H and λ_P are corresponding weights of them. We use these losses to only optimize Θ_H and Θ_P, union of the parameters of hypothesis-only branch and premise-only branch. Note that we do not back propagate these two losses to the encoder, preventing it from directly learning the biases.

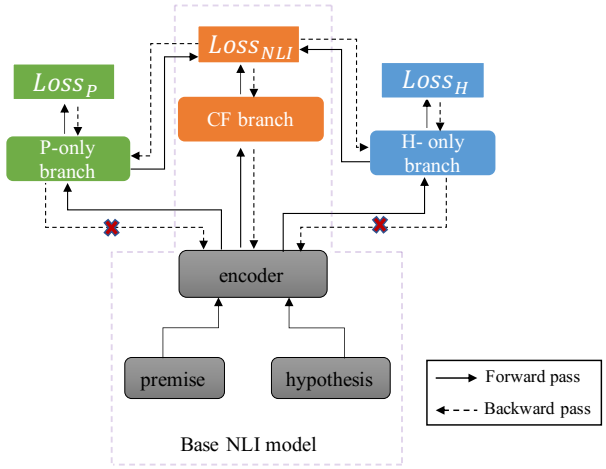


Figure 5: An illustration of the training process.

In the testing stage, we use the debiased effect for inference, which is implemented as:

$$\begin{aligned} TIE &= TE - NDE = S_{h,p,c} - S_{h,p^*,c^*} \\ &= \mathcal{F}(S_h, S_p, S_c) - \mathcal{F}(S_h, S_{p^*}, S_{c^*}) \end{aligned} \tag{15}$$

Evaluation

The experiments are conducted for NLI and fact verification. To show the generality, we use the off-the-shelf uncased BERT (Devlin et al. 2019) implementation of (Wolf et al. 2019), fine-tuning for each task as our main base model.

Datasets

For natural language inference, we train models on the SNLI dataset (Bowman et al. 2015), which is known to contain significant annotation artifacts. The dataset consists of pairs of premise and hypothesis sentences along with their inference labels. We evaluate the models on SNLI-hard (Gururangan et al. 2018), a subset of SNLI test set where a hypothesis-only model cannot correctly predict the labels.

For fact verification, we use the training dataset provided by the FEVER challenge (Thorne et al. 2018). The task concerns about assessing the validity of a claim sentence in the context of a given evidence sentence, which can be labeled as either *support*, *refutes*, and *not enough information*. Schuster et al. (2019) introduced a new evaluation set fever-symmetric dataset to avoid the idiosyncrasies observed in the claims of this benchmark, *i.e.*, the occurrence of words and phrases in the claim that are biased toward certain labels. The collected dataset is challenging, and the performance of the models relying on biases evaluated on this dataset drops significantly. We evaluate the models on the both versions (version 1 and 2) of their test sets [4].

Implementation

Encoder We consider BERT as the base encoder for both tasks, which has shown impressive performance on this task. We fine-tune all models using BERT for 3 epochs and use the default parameters and default learning rate of $1e-5$.

Combined Feature Branch Following the standard setup for sentence pair classification tasks, the predictions of combined feature branch are based on the concatenation of the premise/evidence and the hypothesis/claim with a delimiter token based on the base BERT model.

Premise/Evidence-only Branch The premise/evidence-only model predicts the labels using only premises/evidences as input, which is a shallow nonlinear classifier with 768, 384 and 192 hidden units with $Tanh$ nonlinearity, consistent with the baseline (Mahabadi, Belinkov, and Henderson 2020).

Hypothesis/Claim-only Branch The hypothesis/claim-only model predicts the labels using only hypotheses/claims as input, which is the same shallow nonlinear classifier with premise/evidence-only branch. Besides, because of the existence of the confounder, we use the backdoor adjustment criteria to calculate the causal effect of hypothesis/claim-only branch following Eq. 9 to 11.

Results

We compare our CICR models with state-of-the-art methods:

- **BERT** (Devlin et al. 2019) is the off-the-shelf uncased BERT based model with cross entropy loss.
- **RUBi** (Cadène et al. 2019) is a recently proposed language-prior based methods to alleviate uni-modal biases learned by visual question answering models.
- **DFL** (debiased focal loss) and **PoE** (product-of-experts) (Mahabadi, Belinkov, and Henderson 2020) are two techniques to reduce biases learned by neural models with model ensemble.

For fact verification, we further compare against the Reweight method (Schuster et al. 2019) and Self-debiasing method (Ghaddar et al. 2021).

- **Reweight** introduces a regularization method which alleviates the effect of bias based on the correlation of the n-grams within the claim sentences with the target labels.
- **Self-debiasing** designs a debiasing framework whereby the shallow representations of the main model are used to derive a bias model.

[4] https://github.com/TalSchuster/FeverSymmetric

We omit some other debiasing approaches such as learned-mixin (Clark, Yatskar, and Zettlemoyer 2019), Belinkov et al. (2019), Regularized-conf (Utama, Moosavi, and Gurevych 2020) and LTGR (Du et al. 2021) here and only report the state-of-the-art models due to the space limitation.

Table 2 [5] shows the experimental results on SNLI test set and hard set. Our proposed CICR$_{FC}$ and CICR$_{SUM}$ are highly effective, resulting in 4.11 and 5.3 points gain compared with the BERT-based model in hard set respectively. Besides, we significantly surpass the prior debiasing works of RUBi, DFL and PoE, setting a new state-of-the-art.

Loss	Test	Hard	Δ
BERT	90.53	80.53	-
RUBi	90.69	80.62	+0.09
DFL	89.57	83.01	+2.48
PoE	90.11	82.15	+1.62
CICR$_{FC}$	90.12	84.64	+4.11
CICR$_{SUM}$	90.14	**85.83**	**+5.3**

Table 2: Results on SNLI and hard set.

Table 3 shows the evaluation results on FEVER development and symmetric test sets. From Table 3, we observe that: 1) Debiasing methods outperform BERT-based model on symmetric test sets by large margins. Among these methods, our proposed CICR models achieve the strongest performances on both symmetric test set v1 and v2. 2) Our CICR minimizes the trade-off between the in-distribution and out-of-distribution performance compared to the other methods. For example, RUBi maintains the in-distribution performance but only improves the average accuracy from 56.49% to 57.6% on symmetric test set v1. Reweight and DFL gain 5.11 and 7.53 points improvement over the BERT baseline while reducing the dev accuracy by 1.39 and 2.92 points respectively on symmetric test set v1. Differently, our method achieves the competitive 13.52 and 14.95 points gain without dropping the in-distribution performance, indicating the robustness of counterfactual inference framework.

Loss	Dev	Symmetric Test Set V1	Symmetric Test Set V2
BERT	85.99	56.49	64.4
RUBi	86.23	$57.60_{+1.11}$	$65.38_{+0.98}$
Reweight	84.60	$61.6_{+5.11}$	$66.5_{+2.1}$
Self-debiasing	86.90	$63.8_{+7.31}$	-
DFL	83.07	$64.02_{+7.53}$	$66.57_{+2.17}$
PoE	86.46	$66.25_{+9.76}$	$69.10_{+4.7}$
CICR$_{FC}$	86.08	$70.01_{+13.52}$	$\mathbf{73.45}_{+9.05}$
CICR$_{SUM}$	86.43	$\mathbf{71.44}_{+14.95}$	$72.17_{+7.77}$

Table 3: Results on FEVER and symmetric test set.

[5] Δ column of Table 2 and the $+$ sign of Table 3-4 represent the absolute improvements compared with baseline BERT model.

Effectiveness of Causal Intervention and Counterfactual Reasoning

To investigate the effectiveness of our proposed components of the method, we also perform the ablation experiments. Specifically, when we discard the causal intervention part (w/o CI in Table 4) which means using the traditional likelihood to compute the hypothesis/claim branch, the performance drops, demonstrating the effectiveness of the causal intervention. Besides, when we remove the counterfactual reasoning part (w/o CR in Table 4), the performance has decreased more obviously. The result is reasonable since causal intervention eliminates the influence of spurious correlations in the training stage, while counterfactual reasoning conducts debiasing in the inference stage.

NLI	Test	Hard	Δ
CICR$_{FC}$	90.12	84.64	+4.11
w/o CI	90.17	83.72	+3.19
w/o CR	90.20	82.80	+2.27
CICR$_{SUM}$	90.14	85.83	+5.3
w/o CI	90.44	84.33	+3.8
w/o CR	91.12	83.42	+2.89
Fact Verification	**Dev**	**Symmetric Test Set V1**	**Symmetric Test Set V2**
CICR$_{FC}$	86.08	$70.01_{+13.52}$	$73.45_{+9.05}$
w/o CI	86.24	$69.60_{+13.11}$	$72.47_{+8.07}$
w/o CR	86.58	$67.73_{+12.24}$	$71.65_{+7.25}$
CICR$_{SUM}$	86.43	$71.44_{+14.95}$	$72.17_{+7.77}$
w/o CI	86.59	$70.43_{+13.94}$	$70.65_{+6.25}$
w/o CR	86.50	$67.52_{+11.03}$	$69.76_{+5.36}$

Table 4: Evaluation results on two tasks for ablation study.

Conclusion

In this paper, we propose a novel bias mitigation strategy to reduce known biases learned by NLU models based on causal inference. The detailed implementation consists of de-confounded training with causal intervention and unbiased inference with counterfactual reasoning, which is effective and agnostic to the base encoder models. Specifically, the bias is formulated as the direct causal effect and we extract the unbiased prediction by subtracting the direct bias effect from the total causal effect. Besides, we used the causal intervention with backdoor adjustments to remove the confounder which is the cause of spurious correlations. Experimental results on two NLU tasks: natural language inference and fact verification demonstrate the effectiveness of our CICR. Future work may include developing a more complex causal graph with external knowledge with our counterfactual inference framework.

Acknowledgments

This work was supported by National Key R&D Program of China (2020AAA0109603), State Key Laboratory of Computer Architecture (ICT,CAS) under Grant No. CARCHA202008 and Institute of Precision Medicine, Tsinghua University.

References

Belinkov, Y.; Poliak, A.; Shieber, S. M.; Durme, B. V.; and Rush, A. M. 2019. Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. In *ACL*, 877–891.

Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, N. R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; and Pal, C. J. 2020. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In *ICLR*.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In Màrquez, L.; Callison-Burch, C.; Su, J.; Pighin, D.; and Marton, Y., eds., *EMNLP*, 632–642. The Association for Computational Linguistics.

Cadène, R.; Dancette, C.; Ben-younes, H.; Cord, M.; and Parikh, D. 2019. RUBi: Reducing Unimodal Biases for Visual Question Answering. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *NeurIPS 2019:*, 839–850.

Choi, S.; Park, H.; Yeo, J.; and Hwang, S. 2020. Less is More: Attention Supervision with Counterfactuals for Text Classification. In *EMNLP*, 6695–6704.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *EMNLP-IJCNLP*, 4067–4080.

Dagan, I.; Glickman, O.; and Magnini, B. 2005. The PASCAL Recognising Textual Entailment Challenge. In *MLCW*, 177–190.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Didelez, V.; and Pigeot, I. 2001. Judea pearl: Causality: Models, reasoning, and inference. *Politische Vierteljahresschrift*, 42(2): 313–315.

Du, M.; Manjunatha, V.; Jain, R.; Deshpande, R.; Dernoncourt, F.; Gu, J.; Sun, T.; and Hu, X. 2021. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models. In *NAACL-HLT*, 915–929.

Feder, A.; Keith, K. A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M. E.; Stewart, B. M.; Veitch, V.; and Yang, D. 2021. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. arXiv:2109.00725.

Ghaddar, A.; Langlais, P.; Rezagholizadeh, M.; and Rashid, A. 2021. End-to-End Self-Debiasing Framework for Robust NLU Training. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, 1923–1929.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL-HLT)*, 107–112.

He, H.; Zha, S.; and Wang, H. 2019. Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. In *EMNLP-IJCNLP*, 132–142.

Huang, P.; Zhang, H.; Jiang, R.; Stanforth, R.; Welbl, J.; Rae, J.; Maini, V.; Yogatama, D.; and Kohli, P. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings, EMNLP*, 65–83.

Huang, W.; Liu, H.; and Bowman, S. R. 2020. Counterfactually-Augmented SNLI Training Data Does Not Yield Better Generalization Than Unaugmented Data. In *Proceedings of the First Workshop on Insights from Negative Results in NLP, Insights 2020*, 82–87.

Kaushik, D.; Hovy, E. H.; and Lipton, Z. C. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *ICLR*.

Keele, L. 2015. The Statistics of Causal Inference: A View from Political Methodology. *Political Analysis*, 23(3): 313–335.

Mahabadi, R. K.; Belinkov, Y.; and Henderson, J. 2020. End-to-End Bias Mitigation by Modelling Biases in Corpora. In *ACL*, 8706–8716.

Mahajan, D.; Tan, C.; and Sharma, A. 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *CoRR*, abs/1912.03277.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *ACL*, 3428–3448.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 4885–4901.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.; and Wen, J. 2020. Counterfactual VQA: A Cause-Effect Look at Language Bias. *CVPR*, abs/2006.04315.

Pearl, J. 2013. Direct and indirect effects. *arXiv preprint arXiv:1301.2300*.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19.

Pearl, J.; et al. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3: 96–146.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Durme, B. V. 2018. Hypothesis Only Baselines in Natural Language Inference. In *SEM@NAACL-HLT*, 180–191.

Qian, C.; Feng, F.; Wen, L.; Ma, C.; and Xie, P. 2021. Counterfactual Inference for Text Classification Debiasing. In *ACL/IJCNLP*, 5434–5445.

Richiardi, L.; Bellocco, R.; and Zugna, D. 2013. Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42(5): 1511–1519.

Roese, N. J. 1997. Counterfactual thinking. *Psychological bulletin*, 121(1): 133.

Rubin, D. B. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58.

Schuster, T.; Shah, D. J.; Yeo, Y. J. S.; Filizzola, D.; Santus, E.; and Barzilay, R. 2019. Towards Debiasing Fact Verification Models. In *EMNLP-IJCNLP*, 3417–3423.

Shah, D.; Schwartz, H. A.; and Hovy, D. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *ACL*, 5248–5264.

Singh, R.; and Sun, L. 2019. De-biased Machine Learning for Compliers. *CoRR*, abs/1909.05244.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958.

Stacey, J.; Minervini, P.; Dubossarsky, H.; Riedel, S.; and Rocktäschel, T. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *EMNLP*, 8281–8291.

Teney, D.; Abbasnejad, E.; and van den Hengel, A. 2020. Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. In *ECCV*, 580–599.

Thorne, J.; Vlachos, A.; Cocarascu, O.; Christodoulopoulos, C.; and Mittal, A. 2018. The Fact Extraction and VERification (FEVER) Shared Task. *CoRR*, abs/1811.10971.

Tsuchiya, M. 2018. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In *LREC*.

Utama, P. A.; Moosavi, N. S.; and Gurevych, I. 2020. Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance. In *ACL*, 8717–8729.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR*.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual Commonsense R-CNN. In *CVPR*, 10757–10767.

Wang, Z.; and Culotta, A. 2021. Robustness to Spurious Correlations in Text Classification via Automatically Generated Counterfactuals. In *AAAI*, 14024–14031.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, abs/1910.03771.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2048–2057.

Yang, L.; Li, J.; Cunningham, P.; Zhang, Y.; Smyth, B.; and Dong, R. 2021. Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis. In *ACL/IJCNLP*, 306–316.

Ye, X.; Nair, R.; and Durrett, G. 2021. Evaluating Explanations for Reading Comprehension with Realistic Counterfactuals. *CoRR*, abs/2104.04515.

Zhang, W.; Lin, H.; Han, X.; and Sun, L. 2021. De-biasing Distantly Supervised Named Entity Recognition via Causal Intervention. In *ACL/IJCNLP 2021*, 4803–4813.