Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Full Length Article

# Label-aware debiased causal reasoning for Natural Language Inference

Kun Zhang [a,b,*], Dacao Zhang [a,b], Le Wu [a,b,c], Richang Hong [a,b,c], Ye Zhao [a,b], Meng Wang [a,b]

[a] *School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, Anhui, China*
[b] *Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Hefei, 230601, Anhui, China*
[c] *Institute of Dataspace, Hefei Comprehensive National Science Center, Hefei, 230009, Anhui, China*

## ARTICLE INFO

*Keywords:*
Natural language inference
Spurious correlations
Debiased reasoning
Causal effect

## ABSTRACT

Recently, researchers have argued that the impressive performance of Natural Language Inference (NLI) models is highly due to the *spurious correlations* existing in training data, which makes models vulnerable and poorly generalized. Some work has made preliminary debiased attempts by developing data-driven interventions or model-level debiased learning. Despite the progress, existing debiased methods either suffered from the high cost of data annotation processing, or required elaborate design to identify biased factors. By conducting detailed investigations and data analysis, we argue that label information can provide meaningful guidance to identify these spurious correlations in training data, which has not been paid enough attention. Thus, we design a novel *Label-aware Debiased Causal Reasoning Network* (*LDCRN*). Specifically, according to the data analysis, we first build a causal graph to describe causal relations and spurious correlations in NLI. Then, we employ an NLI model (e.g., RoBERTa) to calculate total causal effect of input sentences to labels. Meanwhile, we design a novel label-aware biased module to model spurious correlations and calculate their causal effect in a fine-grained manner. The debiasing process is realized by subtracting this causal effect from total causal effect. Finally, extensive experiments over two well-known NLI datasets and multiple human-annotated challenging test sets are conducted to prove the superiority of *LDCRN*. Moreover, we have developed novel challenging test sets based on MultiNLI to facilitate the community.

## 1. Introduction

Existing Natural Language Understanding (NLU) problems can be formulated as text classification tasks, in which Natural Language Inference (NLI) is one of the representative tasks. NLI requires an agent to determine the inference relation from premise sentence to hypothesis sentence (Bowman et al., 2015; Zhang et al., 2019). Enormous work has been designed in this field, such as ESIM (Peters et al., 2018), DRr-Net (Zhang et al., 2019), and SemBERT (Zhang et al., 2020c). They have achieved impressive performance, even surpassing human performances on some metrics.

However, existing studies (Gururangan et al., 2018; Poliak et al., 2018; Naik et al., 2018; McCoy et al., 2019; Shah et al., 2020) argue that current NLI models are overestimated. The impressive performance mainly dues to the dependency of *annotation bias* or *spurious correlation*. As shown in Fig. 1, specific language patterns can be used to identify specific semantic relations (e.g., *negation* in hypothesis sentence is often correlated with a *Contradiction* label). This phenomenon is called language bias (Gururangan et al., 2018). Gururangan et al. (2018) have pointed out that annotators preferred to use these language biases to generate NLI data. Moreover, Naik et al. (2018) have proved that NLI models would be misled by these spurious correlations and incorrectly exploit these language patterns (e.g., negation, antonymy, and word overlap) for inference relation prediction.

To investigate how language bias affects the model capability, we conduct a data analysis on the well-known SNLI (Stanford Natural Language Inference) dataset. As reported in Table 1, we fine-tune three PLMs with SNLI training data, and directly apply them to different test sets with different **input settings**. Here, hard test set removes those samples that can be accurately classified using only hypothesis sentences. From the results, all models have impressive performance when using both premise and hypothesis sentences (*NLI task setting*). When using only one of sentences (*bias settings*) as input, things become different. When using only premise sentences, performances over test and hard test sets are similar to random guessing, which is as expected since NLI requires an agent to determine the relation between **two sentences**. However, when it comes to only hypothesis sentences, the performance over test set is better than it over hard test set, and better than random guessing (i.e., 48.23% v.s. 31.85% on average). This phenomenon demonstrates the misuse of these models for spurious correlations among input data. Meanwhile, Swayamdipta et al. (2020)

[*] Corresponding author at: School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, Anhui, China.
*E-mail address:* zhang1028kun@gmail.com (K. Zhang).

https://doi.org/10.1016/j.aiopen.2024.02.001
Received 6 November 2023; Received in revised form 17 January 2024; Accepted 26 February 2024
Available online 13 March 2024

| Sentence pair | Label | Most commonly used words |
|---|---|---|
| P: Two women are embracing while holding packages.<br>H: **Two woman are holding packages**. | Entailment (E)<br>(Word Overlap) | man, always, need, want, still, could |
| P: Two men on bicycles competing in a race.<br>H: Men are riding bicycles **on the street**. | Neutral (N)<br>(Ambiguity) | people, many, something, would, years |
| P: A woman rock-climbs in a rural area.<br>H: The woman **never** gets any exercise. | Contradiction (C)<br>(Negation) | never, not, nothing, always, ever |

**Fig. 1.** Some NLI examples from SNLI dataset and corresponding commonly used words for specific language bias. E.g., (Negation) denotes the specific language bias type.

**Table 1**
Accuracy (%) of PLMs with different input settings on SNLI test and hard test sets (*Prem.* denotes that only premise sentence is used as the model input). SimCSE_B denotes using BERT as the backbone encoder.

| Model | ALL | | Prem. | | Hypo. | |
|---|---|---|---|---|---|---|
| | Test | Hard | Test | Hard | Test | Hard |
| BERT | 88.6 | 79.5 | 31.9 | 31.1 | **44.7** | 31.2 |
| RoBERTa | 90.0 | 82.5 | 32.1 | 32.7 | **47.0** | 32.9 |
| SimCSE_B | 86.9 | 76.0 | 32.4 | 27.3 | **53.9** | 30.3 |
| SimCSE_R | 88.7 | 82.3 | 32.5 | 29.3 | **47.3** | 33.0 |
| Avg. | 88.55 | 80.08 | 32.23 | 30.10 | **48.23** | 31.85 |

performed a similar analysis to prove that biased/cheating features are used by models to achieve high performance. To this end, we can conclude that these spurious correlations will mislead models to make a shortcut by learning these patterns, resulting in vulnerable and poorly generalized performance.

To remove the negative impact of spurious correlations among input data, plenty of debiasing strategies have been proposed. For example, Liu et al. (2022) produced novel WANLI data via targeted augmentations to remove spurious correlations from the data-level perspective. Kaushik et al. (2020) designed a human-in-the-loop system, where annotators were asked to generate counterfactual samples for input data. Apart from data-level manipulations, model-level causality analyses are also proposed, in which causal inference (Pearl et al., 2016) is one of the promising directions. For example, Niu et al. (2021) focused on language bias in Visual Question Answering (VQA) and treated it as the direct causal effect of questions on answers. Therefore, the debiased learning can be achieved by employing Average Treatment Effect (ATE) to remove direct causal effect from total causal effect in VQA. Qian et al. (2021) proposed a model-agnostic debiasing framework, in which a biased model was trained on original training data, and two counterfactual inputs were designed to achieve bias distillation from the biased model.

Despite the impressive performance, these methods are still far from satisfactory. Typically, these methods achieved debiased learning by directly pre-defining the bias types (e.g., pre-defined document-level bias and word-level bias in the most related work Qian et al. (2021)). These heuristic methods lack flexibility and persuasiveness. By revisiting the data annotation process (Gururangan et al., 2018; Bowman et al., 2015) and conducting detailed data analysis, we argue that labels can provide essential information for debiased learning of NLI models, which has not been given enough attention. Therefore, our primary focus turns into the problem: **How to model biased information correlated with labels and how to remove it from NLI models for debiased learning.**

To this end, we propose to exploit the potential of labels and design a novel model-agnostic *Label-aware Debiased Causal Reasoning Network* (*LDCRN*) from a causal-effect perspective. Specifically, based on data analysis, we pinpoint the spurious correlation between hypothesis sentences and inference relations, and argue that labels are helpful to identify spurious correlations. Thus, we first build a causal graph (Pearl et al., 2000) (Fig. 2(B)) to describe causal connections and spurious correlations in NLI data. Then, we leverage conventional NLI models (e.g., RoBERTa Liu et al., 2019) to realize biased learning from training data. Next, we answer the counterfactual question: *"What would the inference relation be if premise sentence is unavailable?"* by designing a scenario in which only hypothesis sentences can be accessed. Meanwhile, to identify spurious correlations correctly, we propose to treat labels as exogenous variables and design a novel label-aware biased module to measure the spurious correlations. After that, the debiased inference is achieved by removing the causal effect of spurious correlations from the total causal effect learned by biased models. Finally, we conducted extensive experiments over two well-known NLI datasets and multiple challenging test sets to demonstrate the superiority of our proposed *LDCRN*. Moreover, we have developed two novel challenging test sets over the MultiNLI dataset to facilitate the community.[1]

## 2. Related work

This section is organized as follows: we first introduce the related work from two aspects. Then, we conclude the distinction of our work.

**Data-driven Interventions.** Researchers argued that current data-driven NLU models tended to learn more about the language bias rather than adequately learning the intended task (Poliak et al., 2018; Naik et al., 2018; McCoy et al., 2019; Shah et al., 2020; Liu et al., 2020; Xiong et al., 2021; Wei et al., 2022; Wu and Gui, 2022; Qi et al., 2023; Qiang et al., 2023). Thus, plenty of data-driven interventions are proposed to realize debiasing from a data perspective. For example, Tsuchiya (2018) conducted lab experiments to prove that existing NLI corpora have a hidden bias that would mislead NLI models to make predictions based on only hypothesis sentences. Naik et al. (2018) and McCoy et al. (2019) developed challenging test sets containing elaborately designed samples that are inconsistent with language biases. By injecting them into original data, the dependence on spurious patterns of models can be effectively alleviated. Zhang et al. (2020a) designed a new unified cross-datasets benchmark with 14 NLI datasets for trustworthy generalization performance evaluation. Nie et al. (2020) designed an iterative, adversarial human-and-model-in-the-loop solution for NLU dataset collection that addressed benchmark longevity and robustness issues. Liu et al. (2022) produced a WANLI corpus via targeted augmentations to remove the spurious correlations according to observations in Swayamdipta et al. (2020). Wang and Culotta (2020) proposed to train a robust text classifier by augmenting the training data with automatically generated counterfactual data. Schlegel et al. (2020) summarized and analyzed the heuristics and spurious correlations in datasets, as well as the shortcomings of existing sentence-matching methods. Meanwhile, Despite the progress, these data-level solutions still suffer from high-cost problems, such as additional manual annotations.

[1] https://github.com/little1tow/MultiNLI-Hard-Test.

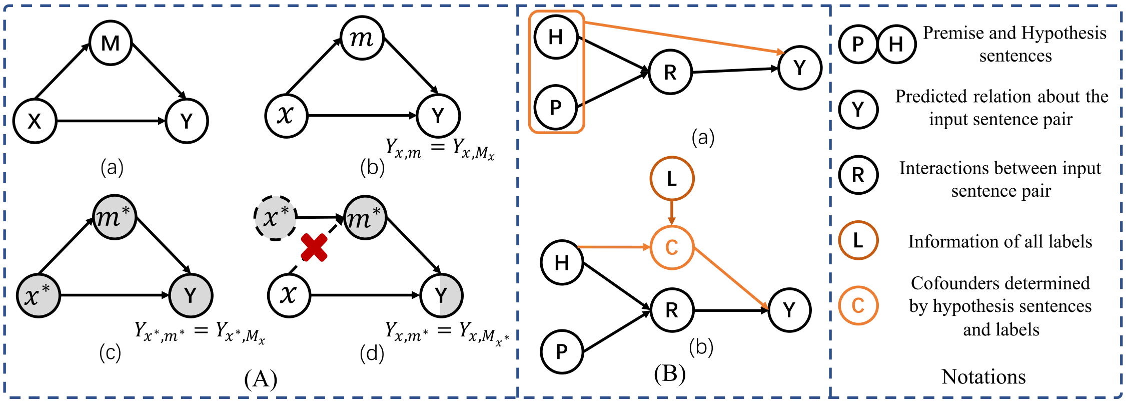

M
X
Y
(a)
$m$
$x$
Y
$Y_{x,m} = Y_{x,M_x}$
(b)
$m^*$
$x^*$
Y
$Y_{x^*,m^*} = Y_{x^*,M_x}$
(c)
$x^*$
$m^*$
$x$
Y
$Y_{x,m^*} = Y_{x,M_{x^*}}$
(d)
(A)
H
P
R
Y
(a)
L
C
H
R
Y
P
(b)
(B)
P H Premise and Hypothesis sentences
Y Predicted relation about the input sentence pair
R Interactions between input sentence pair
L Information of all labels
C Cofounders determined by hypothesis sentences and labels
Notations

**Fig. 2.** (A) Causal graph and counterfactual notations. $X, M$, and $Y$ denote causal, mediator, and effect variables. Gray nodes are at $X = x^*$. (B.a) Commonly used debiased NLI causal graph. (B.b) Our designed causal graph.

**Causal Inference-based Debiasing.** To tackle the shortcoming of data-level solutions, model-level debiasing methods were proposed, such as adversarial training methods (Chai et al., 2022; Qiu et al., 2023) and ensemble-based methods (Karimi Mahabadi et al., 2020; Ghaddar et al., 2021). Among them, integrating causal inference is one representative direction. Causal inference aims at obtaining the causality among treatment (*input*) and effect (*output*) (Pearl et al., 2000, 2016), and has become an important tool in medicine (Konigorski, 2021), neuroscience (Marinescu et al., 2018), cognitive science (Shams and Beierholm, 2010), etc. Recently, researchers have also demonstrated that causal inference can be used to help models remove data biases, improving model robustness and generalization (Utama et al., 2020; Qian et al., 2021; Niu et al., 2021; Ghaddar et al., 2021; Dai et al., 2022; Wu et al., 2022; Sun et al., 2022; Chen et al., 2023).

For example, Niu et al. (2021) focused on visual question answering and developed a counterfactual framework, where language bias was treated as direct causal effect of questions on answers and subtracted from total causal effect for unbiased learning. Choi et al. (2022) proposed identifying causal terms and non-causal terms by calculating the causal "treatment" effect of words on labels, so that better sentence augmentations can be obtained to enhance contrastive learning. Qian et al. (2021) focused on label bias and keyword bias in text classification. They developed a CORSAIR model, where two counterfactual counterparts on inputs were designed to distill and mitigate the biases from a base model. Zhang et al. (2020b) formulated unintended biases in text classification as a kind of selection bias, and proposed to use instance weighting to alleviate the selection bias and constrain the model generalization ability. Gao et al. (2023) incorporated causal inference to analyze the causes of dataset bias. Then, they designed a novel CausalAPM method to project literal and semantic information into independent feature subspaces, and constrain the involvement of literal information in subsequent predictions. Joshi et al. (2022) and Zhou et al. (2023) conducted a detailed analysis about the spurious features in natural language. By reconsidering feature types and training stages, these two works provided a better explanation about the capability of existing debiasing methods and guidance of the debiased learning in natural language. Moreover, Feder et al. (2022) investigated the potential of causal inference to improve the robustness, fairness, and interpretability of NLP models, demonstrating the importance of causal inference. Kıcıman et al. (2023) also pointed out that integrating causal reasoning and LLMs will open a new research frontier.

**Our distinction:** Existing causal inference-based debiasing methods usually pre-defined bias types (Qian et al., 2021) and develop heuristic strategies to realize debiased learning. As a comparison, our method has the following advantages. Firstly, inspired by the works (Joshi et al., 2022; Zhou et al., 2023), we conduct a detailed data analysis (Fig. 1 and Table 1) to better recognize what spurious correlations are and how they affect model performance, which is also the construction basis of our causal graph. Second, we propose that labels can help identify spurious correlations and develop a novel label-aware biased module to conduct fine-grained modeling. Therefore, models learned from observed biased data can be debiased at a minimal cost regarding prediction accuracy reduction. Third, we also design novel challenging test sets based on MultiNLI to facilitate the community.

## 3. *Label-aware Debiased Causal Reasoning Network* (*LDCRN*)

In this section, we first present basic concepts of causal inference. Then, we give technical details of *LDCRN* in a causal effect view.

### 3.1. *Preliminary*

**Causal Graph** is denoted by a Directed Acyclic Graph (DAG): $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ and $\mathcal{E}$ represent the set of variables and causality relations among variables. As shown in Fig. 2(A.a), $X \rightarrow Y$ shows that variable $X$ has *direct* effect on variable $Y$. $X \rightarrow M \rightarrow Y$ is that $X$ has *indirect* effect on $Y$, where $M$ is the mediator. For simplicity, when assigning $X = x$ in Fig. 2(A.b), variable $Y$ can be abbreviated as follows:

$$\begin{aligned} m &= M_x = M(X = x), \\ Y_{x,m} &= Y(X = x, M = m). \end{aligned} \tag{1}$$

**Counterfactual notations** are used to translate causal assumptions from graphs to formulations (Niu et al., 2021). As shown in Fig. 2(A.d), variable $X$ has direct effects on $M$. Therefore, in a factual world, when assigning $X = x$, $M$ will also be affected ($M(X = x)$). In the counterfactual world, $X$ will be simultaneously assigned two different values $x$ and $x^*$. Consequently, variable $Y$ will obtain the counterfactual result $Y_{x,M_{x^*}}$, in which the causal path $X \rightarrow M$ has been blocked.

**Causal effects** are the comparison between two potential outcomes of the same individual given two different treatments (Rubin, 2005). Assuming that $X = x$ is the "treatment" and $X = x^*$ is the "no-treatment", the Total Effect (TE) of treatment $X = x$ on $Y$ can be formulated as follows:

$$TE = Y_{x,M_x} - Y_{x^*,M_{x^*}}. \tag{2}$$

Based on Fig. 2(A.d), TE can be decomposed into Natural Direct Effect (NDE) and Total Indirect Effect (TIE). NDE describes the change of variable $Y$ when $X$ is changed from $x^*$ to $x$ while $M$ is set to the value when $X = x^*$:

$$NDE = Y_{x,M_{x^*}} - Y_{x^*,M_{x^*}}. \tag{3}$$

Then, TIE can be calculated by subtracting NDE from TE as follows:

$$TIE = TE - NDE = Y_{x,M_x} - Y_{x,M_{x^*}}, \tag{4}$$

which denotes the causal effect of $X$ on $Y$ through the mediator $X \rightarrow M \rightarrow Y$. Next, we will introduce the utilization of this framework in NLI and the implementation of each calculation.

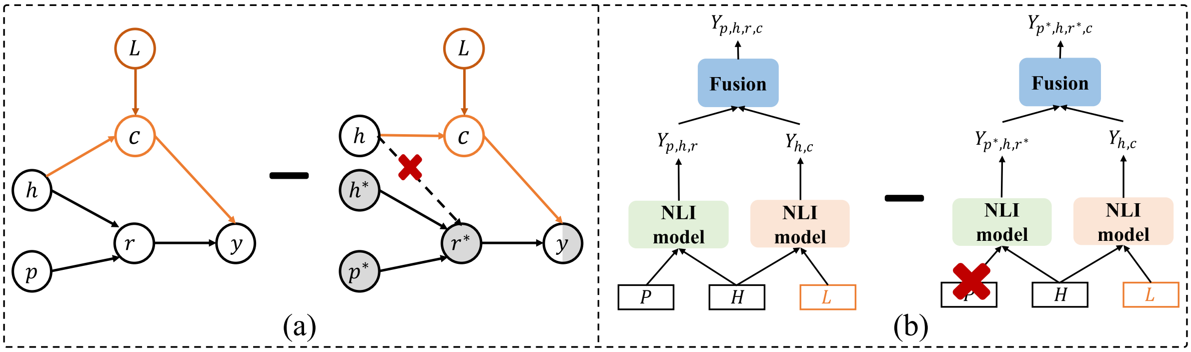


**Fig. 3.** The causal graph and corresponding implementation of our proposed *LDCRN*.

### 3.2. Technical Details of LDCRN

**Causal look of NLI.** Fig. 2(B) illustrates the causal graph (B.a) for traditional debiased NLI models and the causal graph (B.b) for our designed *LDCRN*. In concerned details, input variable $X$ in Fig. 2(A) is transferred to $\{P, H\}$ in Fig. 2(B), representing the input sentence pair $\{S_p, S_h\}$. Mediator variable $M$ in Fig. 2(A) is realized by $R$ and $C$ in Fig. 2(B), where $R$ denotes the interactions between input sentences and $C$ denotes our designed fine-grained mediator for spurious correlations. $L$ in Fig. 2(B.b) represents the all inference relation information. Output variable $Y$ here denotes the predicted inference relation between input sentences $\{P, H\}$. We have to note that existing methods usually treat $L$ only as supervised signals $Y$, and we are concerned more about the identification of labels to spurious correlations among NLI data. Therefore, **compared with existing work, we do not use additional data**.

Meanwhile, in previous debiased NLI models, $\{P, H\} \rightarrow Y$ represents the coarse-grained modeling of spurious correlations, which was used to model all different language biases as one spurious correlation (Poliak et al., 2018; Liu et al., 2020), lacking flexibility and persuasiveness. As a comparison, we argue that the label information will provide guidance when distinguishing the spurious correlations between input sentences and inference relations. Therefore, we use $\{H, L\} \rightarrow C \rightarrow Y$ in Fig. 2(B.b) to describe the fine-grained modeling of *LDCRN*. This strategy is consistent with Fig. 1 and Table 1. For example, in Fig. 1, negation is always connected with contradiction relation. Word overlap is often used to recognize entailment relations. This is also the main difference between our work and existing debiased NLI methods. Next, following the notations in Section 3.1, we can rewrite the causal effect calculation process of NLI models.

According to Eq. (1) and Fig. 2(B.b), when inputs are set as $\boldsymbol{p}$ and $\boldsymbol{h}$, the probability of inference relation can be realized by calculating the effect $Y_{p,h,r,c}$ as:

$$Y_{p,h,r,c} = Y(P = \boldsymbol{p}, H = \boldsymbol{h}, R = \boldsymbol{r}, C = \boldsymbol{c}), \tag{5}$$

where $\boldsymbol{c} = C(L = \boldsymbol{L}, H = \boldsymbol{h})$ is the biased mediator determined by hypothesis sentence $S_h$ and label information $L$. Then, Total Effect (TE) of inputs $X = \{S_p, S_h\}$ in Eq. (2) can be modified as:

$$TE = Y_{p,h,r,c} - Y_{p^*,h^*,r^*,c^*}, \tag{6}$$

where $\{p^*, h^*\}$ denote the no-treatment condition where $\{S_p, S_h\}$ are set to void, $c^* = C(L = \boldsymbol{L}, H = \boldsymbol{h}^*)$, and $r^* = R(P = \boldsymbol{p}^*, H = \boldsymbol{h}^*)$ (see Fig. 3).

In Section 1, we have argued that labels are helpful to model this spurious correlation. Therefore, according to Eq. (3), we intend to calculate the causal effect of spurious correlations from $S_h$ to $Y$ with the consideration of label information $L$, which can be realized by comparing the factual inference (with only $S_h$) and counterfactual inference (even without $S_h$). Since we leverage all label information to guide the modeling of spurious correlations between $H$ and $Y$, and add a biased mediator $C$ to achieve fine-grained modeling, we name this calculation as Controlled Direct Effect (CDE) (Pearl, 2022), replacing the original NDE:

$$CDE = Y_{p^*,h,r^*,c} - Y_{p^*,h^*,r^*,c^*}. \tag{7}$$

Based on Eqs. (6) and (7), we can rewrite Eq. (4) to achieve debiased learning as follows:

$$TIE = TE - CDE = Y_{p,h,r,c} - Y_{p^*,h,r^*,c}, \tag{8}$$

where $TIE$ can be used to achieve the debiased inference relation prediction. To this end, the problem becomes how to calculate $Y_{p,h,r,c}$ and $Y_{p^*,h,r^*,c}$ according to biased observed data.

### 3.3. Implementation of LDCRN

Fig. 3(b) illustrates the implementation details of our proposed *LDCRN*. To realize the debiased learning, it is natural to consider PLMs (e.g., BERT). However, two main problems should be tackled first: (1) *How to realize the counterfactual input (i.e., $\boldsymbol{p}^*, \boldsymbol{r}^*$);* (2) *How to model the biased information contained in labels.* Next, we introduce our designs for these two questions and the implementations of *LDCRN*.

**Q1: Counterfactual Input.** As mentioned before, no-treatment condition is defined as blocking signals from premise sentence $S_p$ (i.e., $S_p = \boldsymbol{\phi}$). Since neural models cannot deal with the inputs that are void, we leverage the average representation $\boldsymbol{a}$ of all premise sentences in training data to replace the void inputs. This intuition makes sense as $\boldsymbol{a}$ can provide data-specific information, which can be treated as the prior knowledge for learning models. One step further, we have compared different settings of $\boldsymbol{a}$ for better illustration in Section 4.

**Q2: Label-aware Biased Module.** We argue that labels can help to model spurious correlations caused by language bias. Therefore, we propose to leverage those words that have strong correlations with labels to measure the biased information contained in labels. Considering that word frequency will select the same words for different labels and lack differentiation, we leverage Pointwise Mutual Information (PMI) to select the most relevant Top-K words for each label as follows:

$$\begin{aligned} PMI(w_j, l_i) &= log \frac{p(w_j, l_i)}{p(w_j, \cdot) \cdot p(\cdot, l_i)}, \\ W_i &= TopK(\{PMI(w_j, l_i)\}), j = \{1, 2, \dots, m\}, \end{aligned} \tag{9}$$

where $p(w_j, l_i)$ denotes the co-occurrence of word $w_j$ and label $l_i$. $m$ is the number of words. Then, we formulate the biased information $\boldsymbol{l}_i$ for $i$th label as the average embedding of corresponding *Top-K* words. After that, we stack biased information of all labels together as the matrix $\boldsymbol{L}$ to denote the biased information from all labels:

$$\boldsymbol{L} = \{\boldsymbol{l}_1, \boldsymbol{l}_2, \dots, \boldsymbol{l}_n\}, \quad \boldsymbol{l}_i = \frac{1}{k} \sum_{w_j \in W_i} PLM(w_j), \tag{10}$$

where $n$ is the number of labels. $PLM(\cdot)$ denotes using PLM (e.g., BERT, RoBERTa) to obtain the embedding of the selected words.

**Table 2**
The statistics of test datasets we used. Note that Hans (McCoy et al., 2019) dataset has two categories (Entailment and Non-Entailment).

| Dataset | E | N | C | All |
|---|---|---|---|---|
| SNLI | 1058 | 1068 | 1135 | 3261 |
| Matched | 1170 | 1154 | 925 | 3249 |
| Mismatched | 1147 | 1077 | 990 | 3214 |
| RP&RH | 508 | 554 | 538 | 1600 |
| Hans | 15,000 | 15,000 | | 30,000 |

**Parameterization.** Similar to previous work (Niu et al., 2021), we leverage two neural models $F_{p,h,r}(\cdot)$ and $F_{h,c}(\cdot)$, as well as one fusion function $G(\cdot)$ to parameterize Eq. (5) as follows:

$$Y_{p,h,r} = F_{p,h,r}(\boldsymbol{p}, \boldsymbol{h}), \quad Y_{h,c} = F_{h,c}(\boldsymbol{h}, \boldsymbol{L}), \\ Y_{p,h,r,c} = G(Y_{p,h,r}, Y_{h,c}), \tag{11}$$

where $Y_{p,h,r}$ describes the interaction path (i.e., $(P,H) \rightarrow R \rightarrow Y$) and $Y_{h,c}$ describes the language bias path (i.e., $(L,H) \rightarrow C \rightarrow Y$). $F(\cdot)$ can be any representation learning model, such as BERT, RoBERTa, and SimCSE.

After obtaining results, we intend to use the combination of $Y_{p,h,r}$ and $Y_{h,c}$ to make the final prediction. Therefore, two different fusion strategies, Concatenation (CON) and SUM, are considered:

$$(CON): G(Y_{p,h,r}, Y_{h,c}) = [Y_{p,h,r}; Y_{h,c}], \\ (SUM): G(Y_{p,h,r}, Y_{h,c}) = Y_{p,h,r} + Y_{h,c}. \tag{12}$$

**Training and Inference.** Following Multi-task learning, we require that each path can make correct predictions. Therefore, a softmax layer is used to process results from each path. *Cross-entropy* is used to optimize learning models. To this end, the optimization target can be formulated as follows:

$$Loss = Loss_1 + \lambda Loss_2 + (1-\lambda) Loss_3, \tag{13}$$

where $\{Loss_1, Loss_2, Loss_3\}$ are *Cross-Entropy loss* over $Y_{p,h,r,c}$, $Y_{p,h,r}$, and $Y_{h,c}$. $\lambda$ is the hyper-parameter to balance the impacts of different components. After finishing model training, we leverage Total Indirect Effect (TIE) for inference relation prediction on test sets and rewrite Eq. (8) as:

$$TIE = TE - DE = Y_{p,h,r,c} - Y_{p^*,h,r^*,c} = G(Y_{p,h,r}, Y_{h,c}) - G(Y_{p^*,h,r^*}, Y_{h,c}). \tag{14}$$

## 4. Experiments

In this section, we first introduce the experimental settings. Next, we report results and detailed analyses of models and results. *Accuracy* (%) on different test sets is selected as the evaluation metric. **Boldface** and underline are used to denote the best and the second-best results.

### 4.1. Experimental settings

**Datasets.** We select SNLI and MultiNLI to verify the model performance. Moreover, multiple challenging test sets are employed to verify the model performance on debiased NLI targets, including SNLI hard (Gururangan et al., 2018), Hans (McCoy et al., 2019), and RP&RH test sets (Kaushik et al., 2020). We also develop two novel challenging test sets based on MultiNLI for debiased NLI performance evaluation. Table 2 summarizes the statistic information of these datasets.

**Challenging test sets construction.** We follow SNLI hard set (Gururangan et al., 2018) to process the MultiNLI data. Specifically, we first train $BERT_{base}$ on MultiNLI data with only the hypothesis sentences. Then, we select the samples with low prediction confidence on the matched and mismatched test sets. The corresponding threshold is set as 0.35. Along this line, those selected samples cannot be correctly predicted with only hypothesis sentences and are used to make up the challenging test sets. The performance of trained $BERT_{base}$ on the original test sets and our constructed challenging test sets is reported in Table 3.

**Table 3**
Performance (Accuracy) of trained $BERT_{base}$ on MultiNLI test sets and challenging test sets.

| | Matched | Mismatched |
|---|---|---|
| Original | 59.5 | 59.5 |
| Challenging | 13.6 | 12.4 |

**Baseline.** To make a comprehensive comparison, we select two types of baselines. (1) Traditional representation learning methods: SimCSE (Gao et al., 2021), SNCSE (Wang et al., 2022); (2) Debiased representation learning methods: Reweight (Clark et al., 2019), Learned-Mixin (LM) (Clark et al., 2019), and CORSAIR (Qian et al., 2021).

**Model Implementation.** We tune hyper-parameters on the official validation set and use *Early Stop* to determine the best values. And we list some common hyper-parameter settings. For the encoder, we select the commonly used pre-trained language models: `BERT-base-uncased` and `RoBERTa-base`. For hyper-parameter settings, the batch size is 64. The learning rate is $lr = 0.00003$. The $\lambda$ in Eq. (13) is set as 0.5. For fusion strategy, we select $CON$ as the final strategy. Additionally, we implement a linear learning rate warm-up for 1000 steps and set the weight decay to 0.05.

### 4.2. Overall experiments

**Debiased NLI performance.** As mentioned in Sections 1 and 3, the debiased learning tries to remove the effect of shortcut usage in the model learning process, which will harm the model performance on the original test set. Therefore, our target is to improve the model robustness (i.e., performance on challenging test sets) at the minimal cost of accuracy loss on the original test sets. With this consideration, we report the overall results on SNLI and MultiNLI datasets in Table 4 and summarize the observations as follows:

Among all debiased models, LM and LM-H achieve impressive debiasing performances on hard test sets, demonstrating their capability of debiased learning. However, their performance on original test sets drops significantly (e.g., an average 4.0% decrease with BERT). In other words, these works focus too much on debiased learning and mistakenly discard some useful features. On the contrary, Reweight and CORSAIR achieve impressive performance on original test sets. However, their performances on hard test sets are not good enough. Since their debiased modules are elaborately designed for pre-defined bias types, these two models have some weaknesses in generalization. Compared with these debiased models, *LDCRN* achieves comparable performance on all hard tests. Moreover, the performance on original test sets shows almost no degradation and in some cases even has an improvement (e.g., performance on MultiNLI matched test set). The phenomenon proves the superiority of *LDCRN*. By building an accurate causal graph and designing a novel label-aware biased module, *LDCRN* can model the spurious correlations in a fine-grained manner, and achieve debiased learning at a minimal cost in terms of prediction accuracy reduction, realizing the target of this work.

Moreover, we employ more challenging test sets (McCoy et al., 2019; Naik et al., 2018) to verify the effectiveness and robustness of *LDCRN* and report results in Tables 5 and 6. From these results, we list our observations as follows:

**Performance on Challenging test sets.** First of all, we can obtain similar performance on baseline methods. LM-H method achieves impressive debiased learning with the cost of a large reduction in accuracy on some simple categories (e.g., *"entailment"* in Han test). Reweight and CORSAIR still have inflexibility and poor generalization in dealing with

**Table 4**
The comparison of **Accuracy (%)** on different NLI test sets. (+2.67%) means the accuracy improvement of *LDCRN* compared with baseline BERT([CLS]) is 2.67%.

| PLMs | Methods | SNLI | | MultiNLI_matched | | MultiNLI_mismatched | |
|---|---|---|---|---|---|---|---|
| | | Test | Hard | Test | Hard | Test | Hard |
| BERT | [CLS] | 88.6 | 79.5 | 83.8 | 70.4 | **84.5** | 71.3 |
| | first-last avg. | 87.6 | 78.2 | 82.5 | 67.7 | 82.4 | 67.6 |
| | +SimCSE | 88.1 | 78.6 | 81.9 | 67.5 | 82.6 | 69.5 |
| | +SNCSE | 88.0 | 79.0 | 82.5 | 68.4 | 82.6 | 68.9 |
| | +Reweight | **88.9(+0.34%)** | 80.8(+1.64%) | **84.2(+0.48%)** | 71.1(+0.99%) | 84.1(−0.47%) | 71.6(+0.42%) |
| | +LM | 84.6(−4.51%) | 82.7(+4.03%) | 81.1(−3.22%) | **77.3(+9.80%)** | 81.1(−4.02%) | **76.9(+7.85%)** |
| | +LM-H | 84.5(−4.63%) | **83.3(+4.78%)** | 80.5(−3.94%) | 76.5(+8.66%) | 80.6(−4.62%) | 76.6(+7.43%) |
| | +CORSAIR | 88.7(+0.11%) | 80.0(+0.63%) | 83.9(+0.12%) | 71.4(+1.42%) | 84.1(−0.47%) | 71.7(+0.56%) |
| | +*LDCRN* | 88.6(+0.00%) | 81.4(+2.39%) | **84.2(+0.48%)** | 72.0(+2.27%) | 83.9(−0.07%) | 71.6(+0.42%) |
| RoBERTa | [CLS] | **90.0** | 82.5 | 87.4 | 76.1 | **87.3** | 75.8 |
| | first-last avg. | 88.5 | 79.8 | 86.0 | 75.0 | 86.0 | 75.0 |
| | +SimCSE | 89.0 | 80.2 | 86.2 | 74.8 | 86.1 | 75.3 |
| | +SNCSE | 89.0 | 80.4 | 86.2 | 74.9 | 86.1 | 75.1 |
| | +Reweight | 89.8(−0.22%) | 82.1(−0.48%) | 87.5(+0.11%) | 76.5(+0.53%) | 87.2(−0.11%) | 76.4(+0.79%) |
| | +LM | 88.7(−1.44%) | 84.4(+2.30%) | 86.8(−0.69%) | **80.7(+6.06%)** | 86.9(−0.46%) | **81.1(+6.99%)** |
| | +LM-H | 88.9(−1.22%) | **84.7(+2.67%)** | 86.8(−0.69%) | 80.5(+5.78%) | 87.0(−0.34%) | **81.1(+6.99%)** |
| | +CORSAIR | 89.5(−0.56%) | 82.8(+0.36%) | 86.2(−1.37%) | 76.0(−0.13%) | 86.4(−1.03%) | 76.3(+0.66%) |
| | +*LDCRN* | 89.8(−0.22%) | 83.4(+1.09%) | **87.6(+0.23%)** | 78.1(+2.63%) | 87.2(−0.11%) | 78.8(+3.96%) |

**Table 5**
**Accuracy** (%) on challenging Hans (McCoy et al., 2019) (2-class classification) and RP&RH (Kaushik et al., 2020) (3-classification) test sets. “NE” denotes non-entailment.

| PLMs | Method | Hans | | | RP&RH | | | |
|---|---|---|---|---|---|---|---|---|
| | | Overall | E | NE | Overall | E | N | C |
| BERT | [CLS] | **57.9** | 96.1 | 19.7 | 70.2 | 75.2 | 64.1 | 71.7 |
| | Reweight | 52.7 | **99.2** | 6.2 | 73.1 | 75.2 | 67.9 | 76.6 |
| | LM-H | 56.5 | 92.0 | **21.1** | **77.8** | **81.3** | **73.1** | **79.2** |
| | CORSAIR | 57.8 | 96.2 | 19.3 | 70.5 | 78.9 | 61.0 | 72.3 |
| | *LDCRN* | 53.5 | 97.4 | 9.6 | 73.1 | 79.9 | 63.5 | 76.4 |
| RoBERTa | [CLS] | 67.3 | **99.0** | 35.7 | 75.3 | 79.1 | 68.2 | 78.6 |
| | Reweight | 71.0 | 98.9 | 43.1 | 74.3 | 78.1 | 68.4 | 76.6 |
| | LM-H | 72.8 | 97.7 | 47.9 | **78.1** | 81.5 | 71.1 | **82.2** |
| | CORSAIR | 69.5 | 98.2 | 40.8 | 75.0 | 76.6 | **71.8** | 76.8 |
| | *LDCRN* | **74.2** | 98.6 | **49.5** | 75.8 | **82.3** | 65.9 | 79.9 |

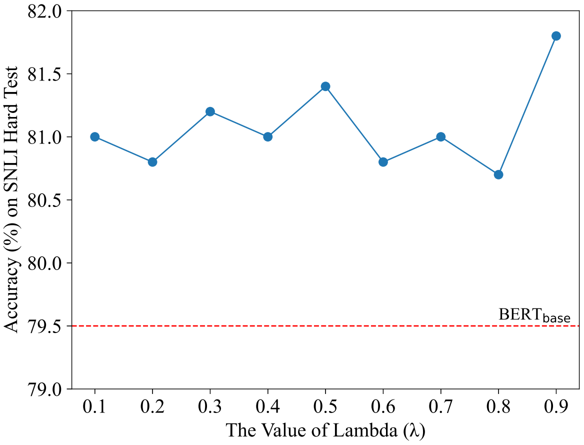


**Fig. 4.** Sensitive test of $\lambda$ on SNLI Hard Test set.

different types of biased information due to their elaborately designed debiased modules.

Meanwhile, *LDCRN* still achieves comparable performance on all test sets with a minimal cost in terms of simple categories accuracy reduction. According to Table 6, we can observe that *LDCRN* enhances the backbone's performance in nearly all aspects. Notably, antonymy, numerical reasoning, and word overlap are the aspects most influenced by our approach. All these results provide strong support for the superiority of our proposed *LDCRN*.

**Performance on Different Backbones.** For backbone encoders, we can observe that using RoBERTa as the backbone achieves better performance than BERT. *LDCRN* also achieves the best performance on Hans test set with RoBERTa. This observation proves that choosing better backbones is a promising direction for model performance improvement. Meanwhile, we also observe that SimCSE and SNCSE do not achieve the expected performance after fine-tuning. One possible reason is that these methods have already been fine-tuned. It is difficult to use additional fine-tuning to further improve the performance, which might even damage the learned parameters.

### *4.3. Ablation study*

The overall experiments have demonstrated the effectiveness of our proposed *LDCRN*. However, it is still unclear which part plays a more critical role in performance improvement. Therefore, we conduct ablation studies on different components to answer this question. Results are summarized in Table 7.

**For fusion strategies**: $CON$ strategy achieves the best performance on hard test set, while $SUM$ strategy has the best performance on the original test set. By comparing with the accuracy of each label, we can obtain that $CON$ strategy leads to a large improvement in entailment and contradiction relations on hard tests. Moreover, $CON$ strategy can keep the learned information as much as possible. These are all helpful for models to distinguish causality from spurious patterns and realize better debiased model learning.

**For the accuracy in each category**: we can observe that models have a relatively small performance decrease on entailment and contradiction relations when conducting evaluations on hard test set. Meanwhile, model performance has a big drop when dealing with neutral relation. Therefore, its accuracy has the most significant impact on overall performance. From the results, our proposed *LDCRN* achieves the least decrease in accuracy on neutral relation, which explains the reason why *LDCRN* can achieve optimal debiased learning.

**For different components**: we compare the modules of different causal effect calculations (i.e., $Loss_3$ for $Y_{h,c}$ and $Loss_2$ for $Y_{p,h,r}$), and the label-aware biased information $L$. According to Table 7(5-7), $Y_{h,c}$ has a bigger impact on the debiased performance. Since $Y_{h,c}$ describes the spurious correlations (i.e., $(L,H) \rightarrow C \rightarrow Y$) in causal graph, removing it will degrade *LDCRN* to a conventional NLI model, limiting the debiasing capability of *LDCRN*. Meanwhile, $L$ describes the biased information associated with all labels. Removing $L$ results in a

**Table 6**
Results on stress test datasets with different backbones.

| Method | Competence test | | | Distraction test | | | | | | Noise test | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Antonymy | | Numerical | Word overlap | | Negation | | Length mismatch | | Spelling error | |
| | Mat | Mis | Reasoning | Mat | Mis | Mat | Mis | Mat | Mis | Mat | Mis |
| BERT | 56.0 | 48.8 | 33.7 | 57.6 | 57.0 | **54.9** | **55.4** | 80.8 | **82.0** | 77.3 | 77.2 |
| +*LDCRN* | **57.3** | **51.2** | **41.3** | **61.5** | **60.6** | 54.7 | 55.1 | **81.6** | 81.8 | **78.0** | **78.3** |
| RoBERTa | 63.7 | 60.7 | 47.0 | 64.2 | 64.2 | 55.4 | 55.7 | 84.3 | **85.2** | 82.0 | 82.0 |
| +*LDCRN* | **67.8** | **62.2** | **51.2** | **69.1** | **68.3** | **56.8** | **56.5** | **85.3** | 84.9 | **83.0** | **83.2** |

**Table 7**
Ablation study of *LDCRN* on SNLI dataset. "w/ SUM" means that changing *CON* fusion strategy to *SUM*. "w/o fusion" means removing the fusion operation.

| Method | SNLI test | | | | SNLI hard test | | | |
|---|---|---|---|---|---|---|---|---|
| | Overall | E | N | C | Overall | E | N | C |
| (1) BERT | 88.6 | 91.5 | 86.3 | 87.9 | 79.5 | 82.8 | 69.6 | 85.7 |
| (2) LDCRN (CON) | 88.6 | 82.4 | 84.0 | 89.1 | **81.4** | **85.8** | 68.9 | 89.1 |
| (3) w/ SUM | 88.7 | 91.8 | 83.9 | **90.2** | 80.7 | 83.9 | 67.0 | **90.5** |
| (4) w/o fusion | 88.6 | 91.3 | **86.9** | 87.5 | 80.4 | 82.7 | 72.6 | 85.6 |
| (5) w/o $\boldsymbol{L}$ | 88.4 | 90.5 | 86.6 | 88.0 | 80.9 | 81.7 | **74.3** | 86.5 |
| (6) w/o $Loss_3$ | 88.7 | 92.2 | 84.8 | 89.0 | 80.5 | 84.8 | 67.7 | 88.6 |
| (7) w/o $Loss_2$ | **88.8** | **92.4** | 86.8 | 88.6 | 81.0 | 85.1 | 69.9 | 87.6 |

large performance decrease on entailment and contradiction relations. These two components (i.e., $Y_{h,c}$, $L$) are both necessary for *LDCRN*. To conduct a detailed analysis, we conducted a parameter sensitive test over $\lambda$ in the following section.

### 4.4. Parameter sensitive test

In the previous text, we have demonstrated that the calculation of $Y_{h,c}$ has a more significant impact on the debiased learning performance. To further analyze how implementations of different causal effect calculations affect the model performance, we conduct additional experiments on hyper-parameters $\lambda$ in Eq. (13), whose values are chosen from $\{0.1, 0.2, 0.3, \ldots, 0.9\}$ and report the corresponding results in Fig. 4.

From the results, we can observe that by incorporating $Y_{h,c}$ and $Y_{p,h,r}$, *LDCRN* achieves impressive improvement on SNLI hard test set, compared with BERT$_{base}$ baseline. Moreover, with the value increase, *LDCRN* becomes unstable. But the overall trend is upward. The best performance is achieved when $\lambda = 0.9$. This behavior can be attributed to the role of $Y_{h,c}$ in controlling the degree of biased learning. While $Y_{h,c}$ is crucial, excessively large values of $Y_{h,c}$ are not conducive to effective biased learning. Therefore, an appropriate value for $\lambda$ enables better learning about bias, ultimately improving the debiasing performance of *LDCRN*.

### 4.5. Case study

To further demonstrate the effectiveness of *LDCRN*, we conduct a qualitative analysis of the model results, which are summarized in Fig. 5. Fig. 5(a)–(c) illustrates some typical language biases. We can observe that BERT exploits language bias to make a shortcut. Thus, it is easily misled by the spurious correlations between specific words and labels, and makes incorrect predictions. For debiased baselines, we can observe that their performance is not stable enough. For example, significant word overlap in sentence pairs will mislead them to make incorrect predictions (Fig. 5(b)). On the contrary, *LDCRN* can successfully measure these biases and rectify them during prediction.

Meanwhile, we also report some bad cases for better illustration. In Fig. 5(d), *LDCRN* overly weakens the connection between negation expression and contradiction label, resulting in an incorrect prediction. This result demonstrates the importance of debiasing intensity in model learning. We should take care of the trade-off between the original target and debiased target. Moreover, Fig. 5(e)–(f) also provide two bad cases. After a detailed analysis, we believe that these cases are more likely to have incorrect ground truth. For example, in Fig. 5(e), the second half of premise has a similar meaning to hypothesis. Therefore, the inference relation should be entailment. From this perspective, *LDCRN* does achieve the impressive debiased target, and demonstrates its effectiveness.

## 5. Conclusion

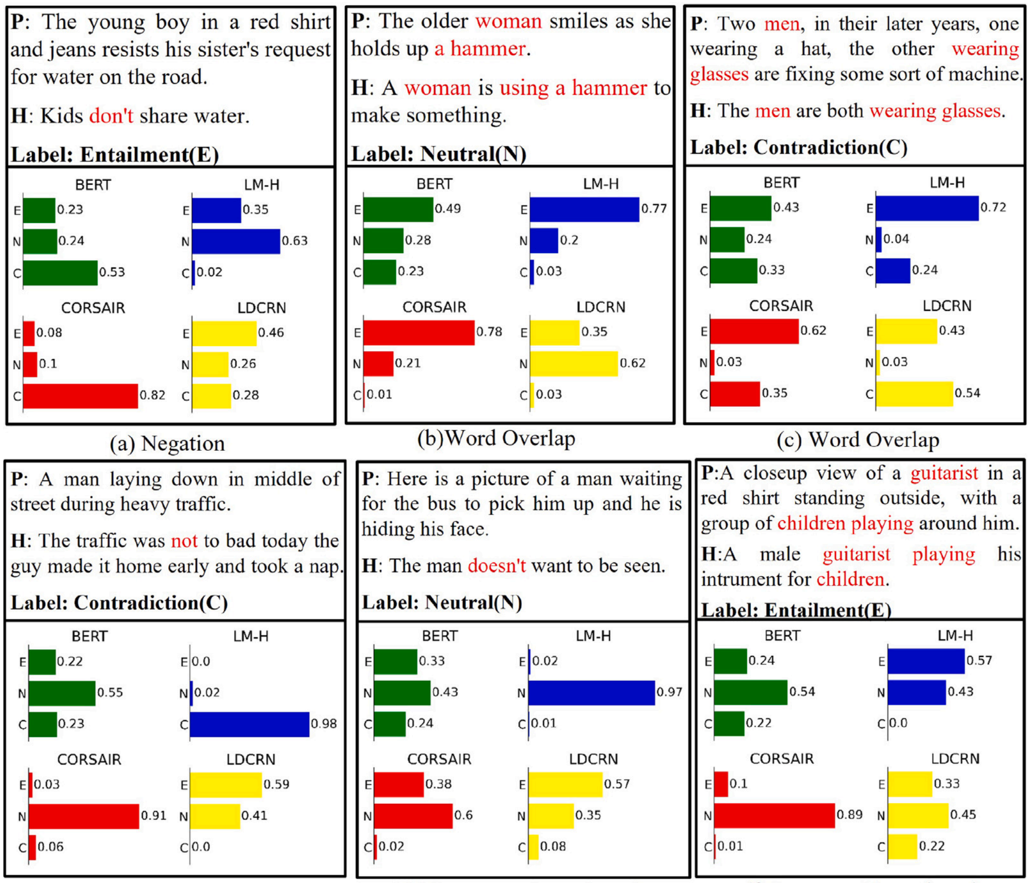
In this paper, we argued that existing debiased NLU models achieved debiased learning by directly pre-defining the bias types, lacking flexibility and persuasiveness. In response, we proposed a novel *LDCRN* to achieve better debiased learning and natural language inference. Specifically, we conducted detailed data analysis to figure out what spurious correlations were and how they were introduced into NLI. Based on the results, we built a causal graph to describe the causal relation and spurious correlations in NLI data. Then, we designed a novel label-aware biased module to realize the fine-grained analysis and employed PLMs to calculate the causal effect of different paths in the causal graph. The debiased learning was achieved by subtracting causal effect of spurious correlations from total causal effect. Extensive experiments on two well-known NLI datasets and multiple challenging hard test sets demonstrated that *LDCRN* could realize impressive debiased performance improvement at a minimal cost in terms of traditional performance reduction. Meanwhile, we constructed two challenge test sets based on MultiNLI to facilitate the community. Since our main contribution is to use label information to guide the fine-grained spurious correlation modeling, in the future, we plan to extend our work to more text classification tasks (e.g., sentiment analysis and paraphrase identification). Moreover, we also plan to leverage the category information of language biases to further enhance debiased learning.

## 6. Limitations

Despite the advantages of pinpointed causal relation and spurious correlation analysis, as well as the designed label-aware biased module, our proposed *LDCRN* still has some space for further improvement. First of all, we leverage Top-K related words to describe biased information, which might not be the best solution for biased information modeling. Taking pre-defined bias definitions and descriptions (Naik et al., 2018; McCoy et al., 2019) might be a better solution. Second, we do not consider the Large Language Model (LLM). How to fully exploit the ability of LLM for spurious correlation recognition as well as conditional data syntactic is also a promising direction. Third, we do not apply our work to different text classification tasks. How to modify our proposed method to measure spurious correlations among different amounts of input sentences and the semantic relations is also one of the promising directions. We will leave these as our future work.

## Acknowledgments

This work is supported by National Science and Technology Major Project (No. 2021ZD0111802), the National Natural Science Foundation of China (No. 62376086, U23B2031, 72188101), Joint Funds of the National Natural Science Foundation of China (No. U22A2094)



(a) Negation (b)Word Overlap (c) Word Overlap

(d) Incorrect prediction (e) ) Incorrect Ground-truth (f) Incorrect Ground-truth

**Fig. 5.** Qualitative comparison of different models on some examples from SNLI hard test set.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 2015. A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 632–642.

Chai, Y., Sun, Z., Qiu, J., Yin, L., Tian, Z., 2022. Tprpf: a preserving framework of privacy relations based on adversarial training for texts in big data. Front. Comput. Sci. 16, 164–618.

Chen, Z., Hu, L., Li, W., Shao, Y., Nie, L., 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 627–638.

Choi, S., Jeong, M., Han, H., Hwang, S.w., 2022. C2l: Causally contrastive learning for robust text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10526–10534.

Clark, C., Yatskar, M., Zettlemoyer, L., 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4069–4082.

Dai, Q., Dong, Z., Chen, X., 2022. Debiased recommendation with neural stratification. AI Open 3, 213–217.

Feder, A., Keith, K.A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M.E., et al., 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. Trans. Assoc. Comput. Linguist. 10, 1138–1158.

Gao, S., Dou, S., Shan, J., Zhang, Q., Huang, X., 2023. Causalapm: Generalizable literal disentanglement for nlu debiasing. arXiv preprint arXiv:2305.02865.

Gao, T., Yao, X., Chen, D., 2021. SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910, URL https://aclanthology.org/2021.emnlp-main.552.

Ghaddar, A., Langlais, P., Rezagholizadeh, M., Rashid, A., 2021. End-to-end self-debiasing framework for robust NLU training. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1923–1929, URL https://aclanthology.org/2021.findings-acl.168.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A., 2018. Annotation artifacts in natural language inference data. In: NAACL-HLT. pp. 107–112.

Joshi, N., Pan, X., He, H., 2022. Are all spurious features in natural language alike? an analysis through a causal lens. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9804–9817, URL https://aclanthology.org/2022.emnlp-main.666.

Karimi Mahabadi, R., Belinkov, Y., Henderson, J., 2020. End-to-end bias mitigation by modelling biases in corpora. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8706–8716, URL https://aclanthology.org/2020.acl-main.769.

Kaushik, D., Hovy, E., Lipton, Z.C., 2020. Learning the difference that makes a difference with counterfactually augmented data. In: International Conference on Learning Representations. ICLR.

Kıcıman, R., Sharma, A., Tan, C., 2023. Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050.

Konigorski, S., 2021. Causal inference in developmental medicine and neurology. Dev. Med. Child Neurol. 63.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, A., Swayamdipta, S., Smith, N.A., Choi, Y., 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 6826–6847, URL https://aclanthology.org/2022.findings-emnlp.508.

Liu, T., Xin, Z., Chang, B., Sui, Z., 2020. HypoNLI: Exploring the artificial patterns of hypothesis-only bias in natural language inference. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 6852–6860, URL https://aclanthology.org/2020.lrec-1.846.

Marinescu, I., Lawlor, P.N., Kording, K.P., 2018. Quasi-experimental causality in neuroscience and behavioural research. Nat. Hum. Behav. 2, 891–898.

McCoy, T., Pavlick, E., Linzen, T., 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. pp. 3428–3448.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., Neubig, G., 2018. Stress test evaluation for natural language inference. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 2340–2353, URL https://aclanthology.org/C18-1198.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., 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. pp. 4885–4901, URL https://aclanthology.org/2020.acl-main.441.

Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R., 2021. Counterfactual vqa: A cause–effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12700–12710.

Pearl, J., 2022. Direct and indirect effects. In: Probabilistic and Causal Inference: The Works of Judea Pearl. pp. 373–392.

Pearl, Judea, Glymour, Madelyn, Jewell, Nicholas P, 2016. Causal Inference in Statistics: A Primer. John Wiley & Sons, New Jersey.

Pearl, J., et al., 2000. Models, Reasoning and Inference. Cambridge University Press, Cambridge, UK, p. 19.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 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). pp. 2227–2237, URL https://aclanthology.org/N18-1202.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B., 2018. Hypothesis only baselines in natural language inference. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 180–191.

Qi, L., Zhang, Y., Liu, T., 2023. Bidirectional transformer with absolute-position aware relative position encoding for encoding sentences. Front. Comput. Sci. 17, 171301.

Qian, C., Feng, F., Wen, L., Ma, C., Xie, P., 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). pp. 5434–5445.

Qiang, Jipeng, Zhang, Feng, Li, Yun, Yuan, Yunhao, Zhu, Yi, Wu, Xindong, 2023. Unsupervised statistical text simplification using pre-trained language modeling for initialization. Frontiers of Computer Science 17 (1), 171–303.

Qiu, H., Feng, R., Hu, R., Yang, X., Lin, S., Tao, Q., Yang, Y., 2023. Learning fair representations via an adversarial framework. AI Open 4, 91–97.

Rubin, D.B., 2005. Bayesian inference for causal effects. Handb. Stat. 25, 1–16.

Schlegel, V., Nenadic, G., Batista-Navarro, R., 2020. Beyond leaderboards: A survey of methods for revealing weaknesses in natural language inference data and models. arXiv preprint arXiv:2005.14709.

Shah, D.S., Schwartz, H.A., Hovy, D., 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5248–5264.

Shams, L., Beierholm, U.R., 2010. Causal inference in perception. Trends in Cognitive Sciences 14, 425–432.

Sun, T., Wang, W., Jing, L., Cui, Y., Song, X., Nie, L., 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 15–23.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N.A., Choi, Y., 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 9275–9293, URL https://aclanthology.org/2020.emnlp-main.746.

Tsuchiya, M., 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC). pp. 1506–1511, URL https://aclanthology.org/L18-1239.

Utama, P.A., Moosavi, N.S., Gurevych, I., 2020. Towards debiasing NLU models from unknown biases. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7597–7610, URL https://aclanthology.org/2020.emnlp-main.613.

Wang, Z., Culotta, A., 2020. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 14024–14031.

Wang, H., Li, Y., Huang, Z., Dou, Y., Kong, L., Shao, J., 2022. Sncse: Contrastive learning for unsupervised sentence embedding with soft negative samples. arXiv preprint arXiv:2201.05979.

Wei, Y., Xue, M., Liu, X., Xu, P., 2022. Data fusing and joint training for learning with noisy labels. Front. Comput. Sci. 16, 166–338.

Wu, T., Gui, T., 2022. Less is better: Recovering intended-feature subspace to robustify NLU models. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 1666–1676, URL https://aclanthology.org/2022.coling-1.143.

Wu, J., Liu, Q., Xu, W., Wu, S., 2022. Bias mitigation for evidence-aware fake news detection by causal intervention. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2308–2313.

Xiong, R., Chen, Y., Pang, L., Cheng, X., Ma, Z.M., Lan, Y., 2021. Uncertainty calibration for ensemble-based debiasing methods. Adv. Neural Inf. Process. Syst. 34, 13657–13669.

Zhang, G., Bai, B., Liang, J., Bai, K., Zhu, C., Zhao, T., 2020a. Reliable evaluations for natural language inference based on a unified cross-dataset benchmark. arXiv preprint arXiv:2010.07676.

Zhang, G., Bai, B., Zhang, J., Bai, K., Zhu, C., Zhao, T., 2020b. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4134–4145, URL https://aclanthology.org/2020.acl-main.380.

Zhang, K., Lv, G., Wang, L., Wu, L., Chen, E., Wu, F., Xie, X., 2019. Drr-net: Dynamic re-read network for sentence semantic matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7442–7449.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., Zhou, X., 2020c. Semantics-aware bert for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9628–9635.

Zhou, F., Mao, Y., Yu, L., Yang, Y., Zhong, T., 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4227–4241.