



# Debiasing NLU models via Causal Intervention and Counterfactual Reasoning

Bing Tian<sup>1</sup>, Yixin Cao<sup>2</sup>, Yong Zhang<sup>1</sup> and Chunxiao Xing<sup>1</sup>

2022/1/27

<sup>1</sup>Tsinghua University, Beijing, China

<sup>2</sup>Nanyang Technological University, Singapore

# Outline



❖ Introduction

❖ Method

❖ Evaluation

❖ Results

❖ Summary

# Introduction

- 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**
- annotation artifacts :
  - *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*
  - *negation is correlated with contradiction label*

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

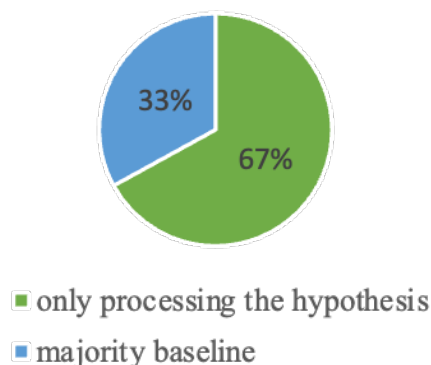
Table 1: Examples from SNLI that illustrate the annotation artifacts.

# Introduction

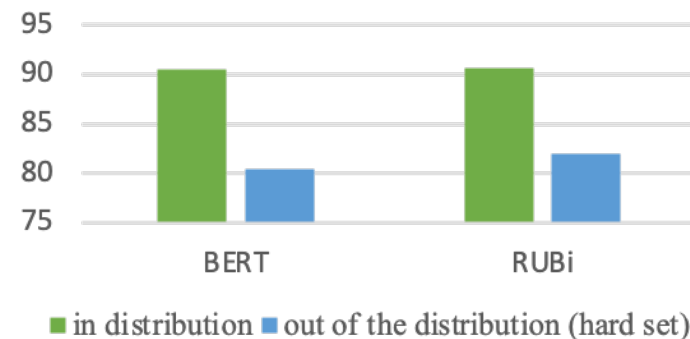
- Natural Language Understanding (NLU) models are prone to relying on idiosyncratic biases (*annotation artifacts*) of the datasets as a *shortcut*, which **goes against the underlying mechanisms of the task of interest**
- annotation artifacts :
  - *the entailed hypotheses tend to replace exact numbers/gender with approximates/generic words (some, at least, human, people etc.)*
  - *purpose clauses are a sign of neutral hypotheses*
  - *negation is correlated with contradiction label*

As a result, with only processing the hypothesis, models can reach accuracy scores as high as twice the majority baseline (67% vs. 33%) when predict the class within the SNLI dataset

Language bias

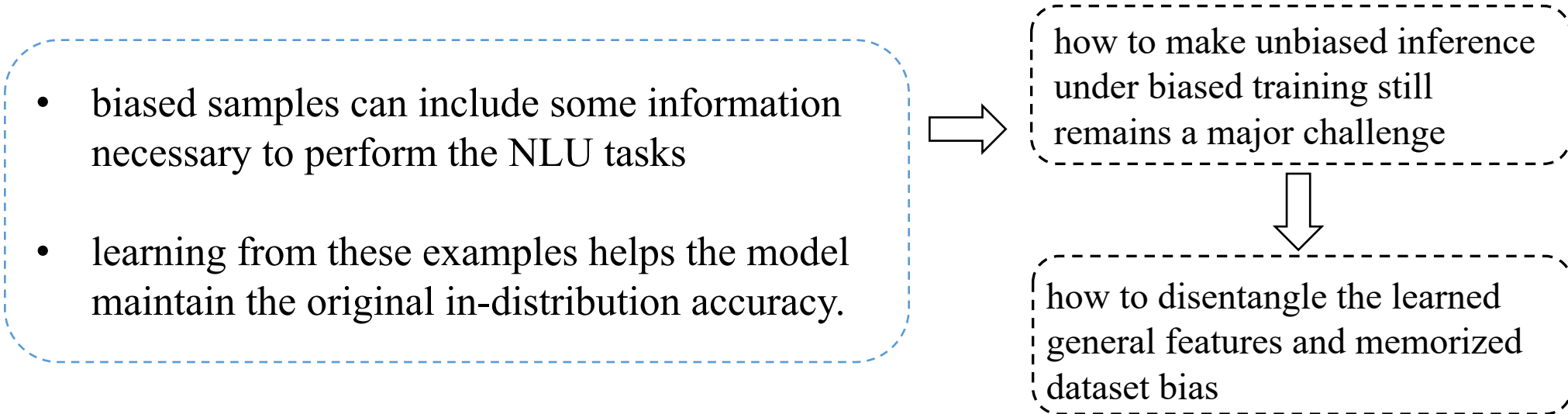


Poor generalization



# Introduction

- Recent popular solution is to develop de-biasing methods that overcome these biases at the training stage
  - first use a bias model to identify biased samples.
  - 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

- 
- biased samples can include some information necessary to perform the NLU tasks
  - learning from these examples helps the model maintain the original in-distribution accuracy.

how to make unbiased inference under biased training still remains a major challenge

how to disentangle the learned general features and memorized dataset bias

# Method

- propose a novel bias mitigation strategy from a causal-effect look
- 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*.
- How could we estimate the causal effects in NLU tasks ?
  - apply de- confounded training with causal intervention to obtain the true causal effects
  - counterfactual reasoning

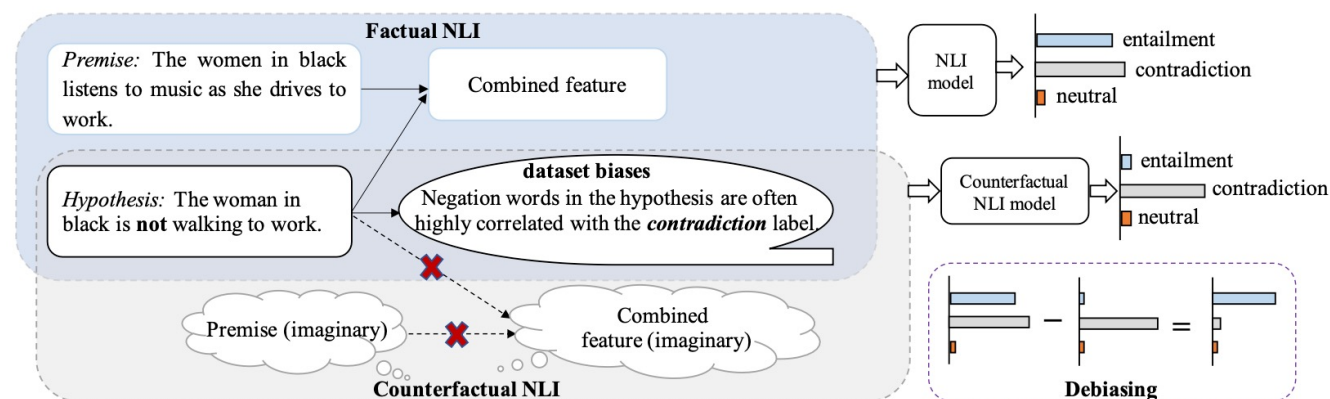
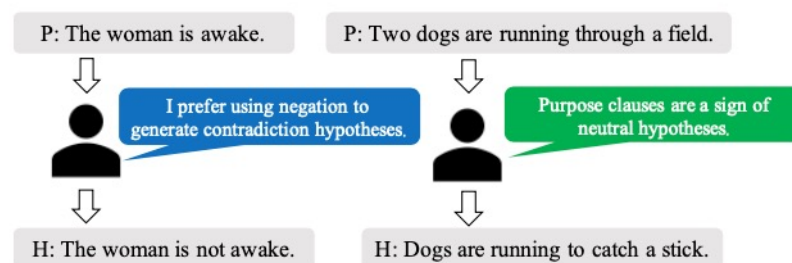


Figure 1: An illustration of factual and counterfactual NLI, as well as the debiasing strategy. Factual NLI depicts the fact where model sees the hypothesis and extracts the combined feature of premise and hypothesis. Counterfactual NLI means that model sees the hypothesis but the combined feature and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



# Method

- De-confounded Training



- counterfactual reasoning
  - two situations

factual NLI: obtain the total causal effects

- both P and H are available
- estimate the total causal effect of P and H on L

counterfactual NLI: obtain the direct causal effect

- “What will the prediction be if seeing the hypothesis sentence only and had not seen the premise and the combined feature?”

# Method

- Causal view

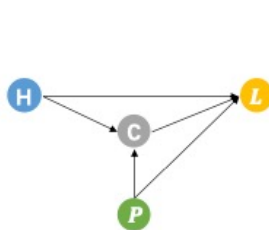
- $TE = S(H = h, P = p, C = c) - S(H = h^*, P = p^*, C = c^*)$

$$= S_{h,p,c} - S_{h^*,p^*,c^*}$$

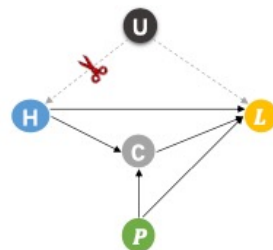
- $NDE = S(H = h, P = p^*, C = c^*) - S(H = h^*, P = p^*, C = c^*)$

$$= S_{h,p^*,c^*} - S_{h^*,p^*,c^*}$$

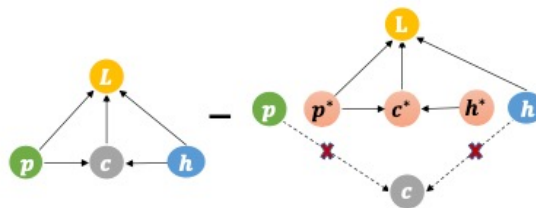
- $TIE = TE - NDE = S_{h,p,c} - S_{h,p^*,c^*}$



(a) causal graph



(b) causal graph with confounder



factual NLI

counterfactual NLI

(c) counterfactual reasoning



# Method

- Parameterization
  - each branch of causal graph can be formulated as a neural model.
  - the score  $S_{h,p,c}$  is calculated through model ensemble with a fusion function

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

- two fusion variants

$$\begin{aligned} \mathcal{F}(S_h, S_p, S_c) &= \mathbf{W}_c S_c + \mathbf{W}_p S_p + S_h \\ \begin{cases} \mathcal{F}(S_h, S_p, S_c) = \log \sigma(S_{SUM}) \\ S_{SUM} = S_h + S_p + S_c \end{cases} \end{aligned}$$

- De-confounder Process

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

# Method

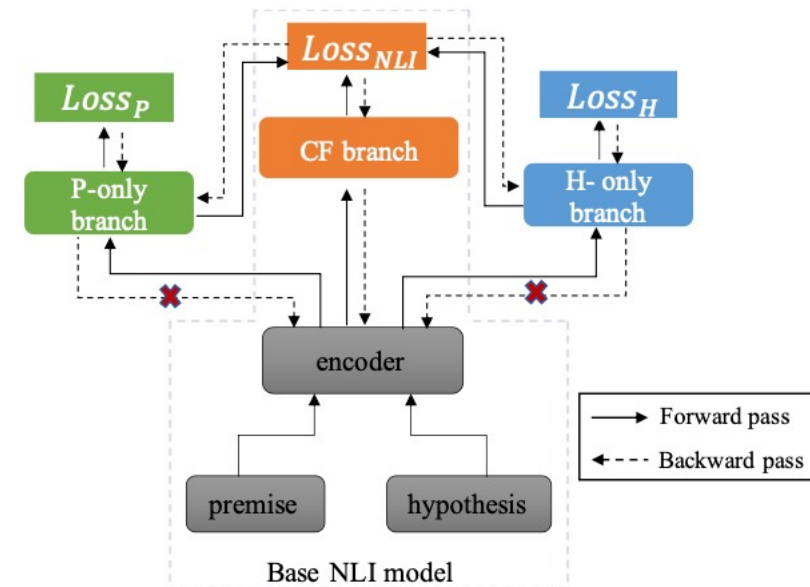
- Parameterization
  - Unbiased Inference
    - In the counterfactual scenery, since the neural models cannot deal with void input, we define the outcome of void input as the *same constant*  $a$  which is a learnable parameter for all the logits.
    - We select the answer with the maximum TIE for inference, which is totally different from traditional strategies that is based on the posterior probability i.e.,  $P(l | h, p)$ .

# Method

- Parameterization
  - Training and Inference
    - Training: 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.
  - Inference: use the debiased effect for inference

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

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



An illustration of the training process

# Evaluation

- Datasets
  - natural language inference
    - SNLI dataset
    - SNLI-hard
  - fact verification
    - Symmetric evaluation set based on the FEVER (fact verification) dataset
- Implementation
  - Encoder
    - BERT as the base encoder for both tasks

# Evaluation

- Baselines
  - BERT: the off-the-shelf uncased BERT based model with cross entropy loss.
  - RUBi : language-prior based methods to alleviate uni-modal biases learned by visual question answering models
  - DFL and PoE : reduce biases learned by neural models with model ensemble
  - Fact verification
    - Reweight : introduces a regularization method
    - Self-debiasing: the shallow representations of the main model are used to derive a bias model

## Results

- $CICR_{FC}$  and  $CICR_{SUM}$  obtain 4.11 and 5.3 points gain compared with the BERT-based model in hard set respectively.
- $CICR_{FC}$  and  $CICR_{SUM}$  significantly surpass the prior debiasing works, setting a new state-of-the-art.
- Our proposed CICR models achieve the strongest performances on both symmetric test set v1 and v2.
- Our CICR minimizes the trade-off between the in-distribution and out-of-distribution performance compared to the other methods.

Loss	Test	Hard	$\Delta$
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	<b>85.83</b>	<b>+5.3</b>

Table 2: Results on SNLI and hard set.

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

Table 3: Results on FEVER and symmetric test set.

# Results

- When we discard the causal intervention part (w/o CI in Table 4), the performance drops, demonstrating the effectiveness of the causal intervention.
- When we remove the counterfactual reasoning part (w/o CR in Table 4), the performance has decreased more obviously.

<b>NLI</b>	<b>Test</b>	<b>Hard</b>	<b><math>\Delta</math></b>
$CICR_{FC}$	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
<b>Fact Verification</b>	<b>Dev</b>	<b>Symmetric Test Set V1</b>	<b>Symmetric Test Set V2</b>
$CICR_{FC}$	86.08	70.01 <sub>+13.52</sub>	73.45 <sub>+9.05</sub>
w/o CI	86.24	69.60 <sub>+13.11</sub>	72.47 <sub>+8.07</sub>
w/o CR	86.58	67.73 <sub>+12.24</sub>	71.65 <sub>+7.25</sub>
$CICR_{SUM}$	86.43	71.44 <sub>+14.95</sub>	72.17 <sub>+7.77</sub>
w/o CI	86.59	70.43 <sub>+13.94</sub>	70.65 <sub>+6.25</sub>
w/o CR	86.50	67.52 <sub>+11.03</sub>	69.76 <sub>+5.36</sub>

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



# Summary

- 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.
- 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.



**Thanks !**