Interventional Training for Out-Of-Distribution Natural Language Understanding

Sicheng Yu[1,2]* **Jing Jiang**[1] **Hao Zhang**[3] **Yulei Niu**[4] **Qianru Sun**[1] **Lidong Bing**[†2]
[1]Singapore Management University [2]DAMO Academy, Alibaba Group
[3]Nanyang Technological University [4]Columbia University
scyu.2018@phdcs.smu.edu.sg, {jingjiang,qianrusun}@smu.edu.sg
hzhang26@outlook.com, yn.yuleiniu@gmail.com
l.bing@alibaba-inc.com

Abstract

Out-of-distribution (OOD) settings are used to measure a model's performance when the distribution of the test data is different from that of the training data. NLU models are known to suffer in OOD settings (Utama et al., 2020b). We study this issue from the perspective of causality, which sees *confounding bias* as the reason for models to learn spurious correlations. While a common solution is to perform intervention, existing methods handle only known and single confounder (Pearl and Mackenzie, 2018), but in many NLU tasks the confounders can be both unknown and multifactorial. In this paper, we propose a novel interventional training method called Bottom-up Automatic Intervention (BAI) that performs multi-granular intervention with identified multifactorial confounders. Our experiments on three NLU tasks, namely, natural language inference, fact verification and paraphrase identification, show the effectiveness of BAI for tackling different OOD settings. [1]

1 Introduction

From the era of word embeddings (Pennington et al., 2014) to pre-trained language models (Devlin et al., 2019), researchers of natural language understanding (NLU) have tried to push the performance on benchmark datasets. Traditional settings assume *independent and identical distribution* (IID) in training and testing splits. However, the IID setting cloaks the vulnerability of neural models, *i.e.*, neural models tend to learn non-robust "shortcut" patterns in the training data but fail to make robust predictions on unseen samples. To evaluate the robustness of models, the *out-of-distribution* (OOD) setting draws the attention of the NLU community.

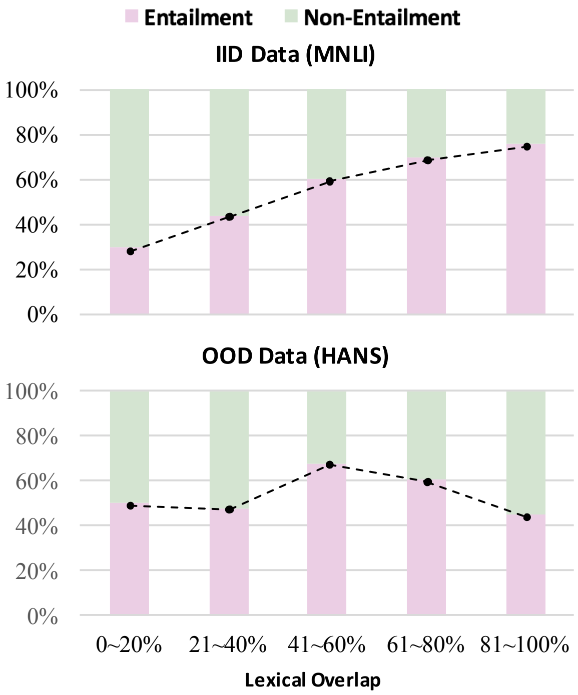


Figure 1: The proportions of entailment and non-entailment samples with different percentages of lexical overlap.

For example, the task of natural language inference (NLI) determines whether a hypothesis can be entailed from a premise. We can observe that the lexical overlap between the hypothesis and the premise correlates with the *entailment* label on the benchmark MNLI dataset (Williams et al., 2018) (as shown in the top part of Figure 1). McCoy et al. (2019) proposed an OOD set named HANS for NLI. As shown in the bottom part of Figure 1, HANS does not have the correlation between lexical overlap and the entailment label. NLI models that rely on the lexical overlap heuristic suffer from a significant degradation on HANS (Utama et al., 2020b).

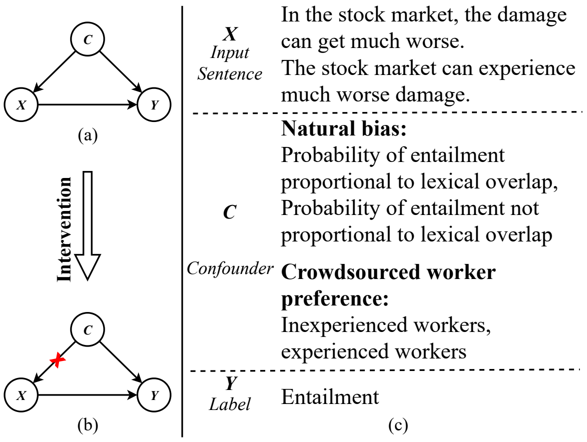
Recently, causal inference has been adopted in NLP to identify robust correlations by analyzing reliable causal effects between variables (Zhang

[*] The work was done when Sicheng was an intern at Alibaba and Yulei was at Nanyang Technological University.

[†] Corresponding author.

[1] Our code is publicly available on GitHub: https://github.com/PluviophileYU/BAI

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11627–11638
December 7-11, 2022 ©2022 Association for Computational Linguistics



In the stock market, the damage can get much worse.
The stock market can experience much worse damage.

Natural bias:
Probability of entailment proportional to lexical overlap,
Probability of entailment not proportional to lexical overlap

Crowdsourced worker preference:
Inexperienced workers, experienced workers

Entailment

Figure 2: (a) Causal graph of NLU tasks, (b) intervention operation, and (c) an example of each node in the causal graph on the NLI task, where the data sample is from MNLI (Williams et al., 2018).

et al., 2021; Nan et al., 2021). From the perspective of causality (Pearl, 2009, 2010), the crux under a model's vulnerability is *confounding bias*. We summarize the causal relations behind NLU tasks as a causal graph in Figure 2(a). X represents the input, *e.g.*, a pair of sentences for NLI, and Y represents a label to be predicted. $X \rightarrow Y$ represents the desired relation for a robust NLU model, *i.e.*, how to predict the label with reliable understanding of the input. $X \leftarrow C \rightarrow Y$ denotes a backdoor path of some unreliable relation between X and Y confounded by the confounder C. Examples of C include nature bias in the dataset (Tang et al., 2020) or crowdsourced workers preference (Geva et al., 2019). For instance, in NLI, C may represent the degree of lexical overlap between the premise and the hypothesis, which is correlated with the entailment relation in the MNLI dataset (see Figure 1).[2] When crowdsourced workers are engaged to create hypotheses for NLI, C could be the experience level of a worker, with inexperienced workers more likely to write simple sentences with straightforward meanings. As a result, these examples of C will make X and Y spuriously correlated.

A common solution of deconfounding is *intervention* (Pearl et al., 2016; Pearl and Mackenzie, 2018), which aims to block the backdoor path (or spurious correlation) by cutting off $C \rightarrow X$ (see Figure 2 (b)). The key idea is to stratify X into different environments (Arjovsky et al., 2019; Teney et al., 2021), *i.e.*, several subsets of training data, according to the identified confounder. Then the model is expected to make environment-agnostic prediction. By doing so, we are controlling X and thus break the backdoor path by D-Separation (Koller and Friedman, 2009). Figure 2(c) depicts an example where the NLI training data is stratified into several environments, *e.g.*, one with obvious trend of lexical overlap bias and another does not. Then the NLI model is trained to fit both environments.

However, the confounder C is not always observed. Furthermore, confounders can be multifactorial in NLU, *e.g.*, it may contain both inherent dataset bias and artifacts from crowdsourced workers. Both scenarios make intervention non-trivial. In this paper, we propose BAI, a bottom-up automatic intervention method, which can (1) identify the unobserved confounder(s) automatically, and (2) perform multi-granular intervention to handle multifactorial confounders. Inspired by Creager et al. (2021), the *automatic* stratifying mechanism is realized by maximizing the difference between data in different environments.[3] We further propose a novel *bottom-up* intervention mechanism that aims to address the multifactorial characteristic of C. While most existing debiasing work only considers a single bias, our bottom-up mechanism enables the model to pick up different confounders in two rounds of interventions. Specifically, based on our preliminary experiments, we find that fine-grained partition (*i.e.*, partition with more environments) results in smaller differences between environments, making environment-agnostic learning easier. Thus we start from a fine-grained partition. We then move on to a coarse-grained partition to further block the backdoor effect via C and make the learning environment-agnostic.

We apply BAI on three OOD benchmarks for NLU tasks. The results show that our method outperforms state-of-the-art methods, *e.g.*, achieving 7 percentage points of absolute gains from the previous best method under OOD setting of Quora Question Pairs (QQP) (Zhang et al., 2019), a benchmark dataset for paraphrase identification.

Contributions: (1) we analyze the issue of NLU vulnerability from the perspective of causality analysis; (2) we propose a bottom-up automatic intervention method to perform intervention for unobserved and multifactorial confounders; and (3)

[2] We highlight that the lexical overlap bias is an example for the purpose of illustration and verification only. Our method is designed for situations with unknown confounders.

[3] Here an environment refers to a subset of training data. A partition is an assignment of the whole training set into multiple environments, *e.g.*, a partition with five environments.

extensive experiments on three OOD benchmarks demonstrate that our method outperforms state-of-the-art methods.

2 Related Work

OOD Generalization. OOD settings have been studied in recent years in NLU. To tackle dataset bias, most existing work relies on instance reweighting with a bias model for debiasing. Specifically, these methods (Cadene et al., 2019) first design a bias model and then train a target debiased model fused with the bias model. Training instances predicted correctly by the bias model will be down-weighted in the training of the debiased model. Early work mainly revolves around different fusion methods (He et al., 2019; Clark et al., 2019; Utama et al., 2020a; Mahabadi et al., 2020) with known bias. Then researchers started looking into unknown bias by designing the bias model with heuristics, *e.g.*, a model trained with very small amount of data (Utama et al., 2020b) or a model with only the bottom layers of the language model (Ghaddar et al., 2021).

However, instance reweighting based methods rely on either prior knowledge of bias or heuristic design of the bias model. Furthermore, it is pointed out that such bias models may not be able to predict the main model's reaction of biased samples and reweighting may waste data (Amirkhani and Pilehvar, 2021). In contrast, our method is derived from causal inference (Pearl and Mackenzie, 2018), which is not related to any form of reweighting. Meanwhile, our method does not adopt any bias model (which requires carefully design or prior knowledge of bias).

Causal Intervention. Causality inference (Pearl, 2009, 2010) measures the causal effect between variables and has been widely applied to various scenarios, *e.g.*, social science (Baron and Kenny, 1986), medical science (Hall et al., 1993), and other applications (Niu et al., 2021; Yu et al., 2020; Niu and Zhang, 2021). Recently, causality inference is introduced to the machine learning community and intervention is one of the techniques in causality inference. Intervention (Pearl, 1993) helps to eliminate the effect of confounders (Yang et al., 2021; Yue et al., 2020; Qi et al., 2020; Nan et al., 2021; Zhu et al., 2022; Niu et al., 2022). Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) implements intervention by learning a model invariant to different environments (Arjovsky et al., 2019; Wang et al., 2021). Although IRM has been widely adopted in computer vision (CV) (Krueger et al., 2021; Rosenfeld et al., 2020; Creager et al., 2021; Liu et al., 2021; Wang et al., 2021; Teney et al., 2021), to the best of our knowledge, our proposed BAI is an initial work of applying IRM in NLU. Our work is distinguishable from previous work in two aspects. First, previous work in CV mainly focuses on image with annotated background as confounder while the confounder in NLU is more abstract and vague. Second, our method is the first work considering multiple partitions for handling multifactorial confounders.

3 Method

3.1 Preliminaries

Causal Intervention is the core idea of this paper. We formulate NLU tasks with a causal graph (Pearl and Mackenzie, 2018), which illustrates the causal relationships between variables with a directed acyclic graph. As shown in Figure 2, each node represents a variable, *e.g.*, a pair of sentences or a label for NLU tasks, and each directed edge denotes that the head node has direct effect on the tail node.

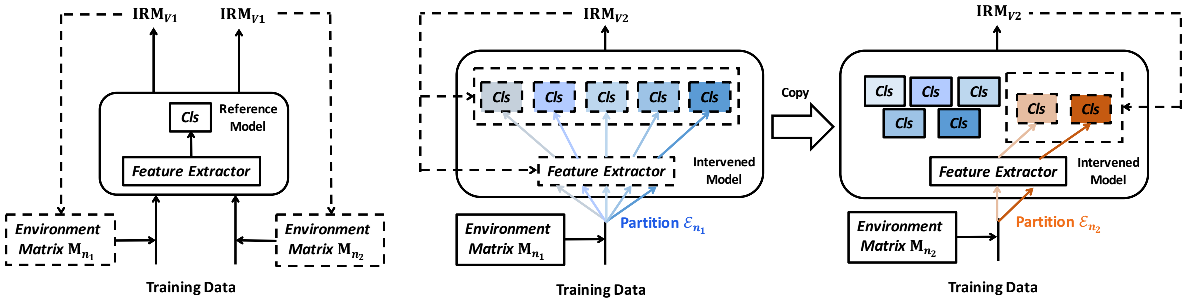
Naïve model training, *i.e.*, empirical risk minimization (ERM) (Vapnik, 1991), indiscriminately learns both spurious correlation $X \leftarrow C \rightarrow Y$ and causal correlation $X \rightarrow Y$. Specifically, by applying Bayes' rule on Figure 2(a), we can obtain:

$$P(Y|X) = \sum_c P(Y|X, c)\underline{P(c|X)}, \qquad (1)$$

where the bias is introduced via $P(C|X)$. For example, consider the NLI task. Let X be a pair of two sentences (premise and hypothesis) and Y the entailment label. Let C represents the degree of lexical overlap between the two sentences in X, and let c_1 and c_2 denote two situations: having obvious lexical overlap and having little or no lexical overlap. Typically on IID training data of NLI, $P(c_1|X)$ is larger than $P(c_2|X)$, and thus $P(c_1|X)$ tends to dominate the overall term, $P(Y|X)$. In other words, model tends to learn $P(Y|X)$ from c_1 instead of X.

In contrast, causal intervention in Figure 2(b) yields:

$$P(Y|do(X)) = \sum_c P(Y|X, c)\underline{P(c)}, \qquad (2)$$



(a) Automatic Stratifying

(b) Bottom-up Intervention

Figure 3: Overall training pipeline of BAI: (a) automatic stratifying where $\mathbf{M}_{n_1}$ and $\mathbf{M}_{n_2}$ are optimized individually; and (b) bottom-up intervention. The dashed arrows denote the back-propagation. Only the modules (or parameter matrices) with dashed box are updated.

where the $do(X)$ denotes that intervention is conducted on X. With do operation, c is no longer associated with X and thus the model treats c_1 and c_2 fairly subject to the prior distribution of C.

Invariant Risk Minimization (Arjovsky et al., 2019) (IRM) is one of the popular tool for intervention in deep neural networks. Given the stratified environments, IRM targets at a robust model which is invariant to environments. In our paper, we utilize two versions of IRM.

Given the input X, model f and the partition of environments $\mathcal{E}$, the original version of IRM (Arjovsky et al., 2019) minimizes the objective:

$$\begin{aligned}\mathrm{IRM}_{v1} = &\sum_{e\in\mathcal{E}} \mathrm{XE}(f(X^e), Y) \\ &+ \lambda \cdot \|\nabla_{\mathbf{w}|\mathbf{w}=\mathbf{1.0}} \mathrm{XE}(\mathbf{w}\cdot f(X^e), Y)\|^2,\end{aligned} \tag{3}$$

where X^e denotes the data in the environment of e and XE denotes cross-entropy loss. $\mathbf{w}$ is a fixed dummy classifier. The second term measures the optimality of $\mathbf{w}$ for each environment to encourage the model to make environment-invariant predictions. This version of IRM is unstable due to the second-order derivatives.

Another version of IRM (Teney et al., 2021) adopted in our paper initializes individual classifier $\mathbf{W}_e$ for each environment e while all environments share one feature extractor. Here we denote the model for the environment e as $f^e = \mathbf{W}_e \circ \Phi$ where Φ is a feature extracter, e.g., BERT. The corresponding loss is written as:

$$\mathrm{IRM}_{v2} = \sum_{e\in\mathcal{E}} \mathrm{XE}(f^e(X^e), Y) + \lambda \cdot \underset{e'\in\mathcal{E}}{\mathrm{Var}}(\mathbf{W}_{e'}). \tag{4}$$

The second term is the variance of classifier weights, which encourages optimal classifiers for different environments to be close to each other[4].

3.2 Bottom-up Automatic Intervention

To implement intervention on NLU tasks with the unobserved and multi-factorial confounder, we propose a Bottom-up Automatic Intervention (BAI) method using IRM. Figure 3 and Algorithm 1 (in Appendix) show the overall pipeline of BAI. It consists of two components: *automatic stratification* and *bottom-up intervention*. The automatic stratification component generates partition of environments by maximizing the difference between data in different environments based on a reference model. The bottom-up intervention component performs intervention at two levels of granularity.

Automatic Stratification generates the partition of environments with unobserved confounder. A good partition is achieved when a reference model behaves differently under different environments. Inspired by Creager et al. (2021), we first train a reference model f_{ref} through the naïve trained BERT (Devlin et al., 2019). Note the second term of Eq. 3 is to make environment-invariant prediction, that is, to minimize the difference of data behavior across environments. Inversely, our goal is to maximize the difference of data behavior by magnifying the second term of IRM.

As shown in Figure 3(a), we initialize an environment matrix $\mathbf{M} \in \mathbb{R}^{D\times N}$ indicating the belonging of each training sample to each environment, where D and N denote the number of training data and pre-defined environments, respectively. $\mathbf{M}^{i,j}$ is the probability of i-th sample belonging to j-th environment. IRM$_{v2}$ is not applicable since the naïve trained reference model only has one classifier. Thus we derive $\mathbf{M}$ by fixing the refer-

[4] More details about IRM in Appendix Section **??**.

ence model f_{ref} and maximizing the second term of IRM$_{v1}$ as follows:

$$\max_{\mathbf{M}} \sum_{e\in\mathcal{E}} \|\nabla_{\mathbf{w}|\mathbf{w}=\mathbf{1.0}}\text{XE}(\mathbf{w}\cdot f_{\text{ref}}(X^e), Y)\|^2, \quad (5)$$

where $\mathcal{E}$ is the partition of environments determined by $\mathbf{M}$. Note that $\max$ operation makes the back-propagation of gradients from $\mathbf{M}$ infeasible. To address this issue, we deploy the Gumbel Softmax trick (Jang et al., 2016) to re-formulate the discrete sampling as:

$$\mathcal{E} = g(\mathbf{M}) = \text{Gumbel-Softmax}(\mathbf{M}). \quad (6)$$

We term the environment matrix with n environments as $\mathbf{M}_n$. Specifically, we deploy automatic stratifying to extract two environments matrices, *i.e.*, fine-grained $\mathbf{M}_{n_1}$ and coarse-grained $\mathbf{M}_{n_2}$ ($n_1 > n_2$), for bottom-up intervention.

Bottom-Up Intervention adopts multi-granular partitions for intervention in a bottom-up fashion, to derive a robust model f_{int}. As shown in Figure 3(b), bottom-up intervention consists of two rounds of intervention deployed by IRM$_{v2}$ due to its stability and scalability.

We first generate fine-grained partition $\mathcal{E}_{n_1}$ and coarse-grained partition $\mathcal{E}_{n_2}$ from $\mathbf{M}_{n_1}$ and $\mathbf{M}_{n_2}$ (see Figure 3(b)), where the number of environments in $\mathcal{E}_{n_1}$ is larger than that in $\mathcal{E}_{n_2}$. Second, we start from the fine-grained partition $\mathcal{E}_{n_1}$ and train the intervened robust model f_{int}. Similarly, we decompose $f_{\text{int}} = \mathbf{W} \circ \Phi$ where Φ is feature extractor, *e.g.*, BERT, and $\mathbf{W}$ is a set of learned classifiers. We use $\mathbf{W}_e$ to represent the classifier exclusive to environment e and $\mathbf{W}\{\mathcal{E}\}$ to denote the set of classifiers for $\mathcal{E}$ partition, that is, $\mathbf{W}\{\mathcal{E}_{n_1}\} = \{\mathbf{W}_e \mid e \in \mathcal{E}_{n_1}\}$ represents all classifiers for partition $\mathcal{E}_{n_1}$. The feature extractor and the classifiers of $\mathcal{E}_{n_1}$ in bottom fine-grained intervention are optimized by:

$$\min_{\Phi,\mathbf{W}\{\mathcal{E}_{n_1}\}} \sum_{e\in\mathcal{E}_{n_1}} \text{XE}(f^e_{\text{int}}(X^e), Y) + \lambda \cdot \operatorname*{Var}_{e'\in\mathcal{E}_{n_1}} (\mathbf{W}_{e'}), \quad (7)$$

Then we conduct the intervention of coarse-grained partition $\mathcal{E}_{n_2}$. To prevent the catastrophic forgetting, *i.e.*, the intervention with new partition may make the model forget the invariant property on previous partition, we incorporate the idea from continual learning (Li and Hoiem, 2017; Rebuffi et al., 2017). Specifically, we fix the parameter of model f_{int} including the feature extractor and n_1 classifiers for $\mathcal{E}_{n_1}$. Then we augment n_2 classifiers for the new partition $\mathcal{E}_{n_2}$, resulting in $n_1 + n_2$ classifiers. Here we only optimize the n_2 augmented classifiers during training as:

$$\min_{\mathbf{W}\{\mathcal{E}_{n_2}\}} \sum_{e\in\mathcal{E}_{n_2}} \text{XE}(f^e_{\text{int}}(X^e), Y) + \lambda \cdot \operatorname*{Var}_{e'\in\mathcal{E}_{n_1}\cup\mathcal{E}_{n_2}} (\mathbf{W}_{e'}), \quad (8)$$

where the first term is based on the new partition $\mathcal{E}_{n_2}$ while the second term computes the variance of classifier weights across all $n_1 + n_2$ classifiers.

Inference is based on the design of IRM$_{v2}$ (Teney et al., 2021). Since we are not able to distinguish which environment the input data belongs to, we simply average the weight of $n_1 + n_2$ classifiers for inference:

$$\hat{Y} = f^{\bar{e}}_{\text{int}}(X) = \bar{\mathbf{W}} \cdot \Phi(X), \quad (9)$$

where $\bar{\mathbf{W}}$ denotes the mean weight of all classifiers.

The overall pipeline of BAI is summarized in Algorithm 1.

Algorithm 1 BAI Training

```
Input: Dataset $\mathcal{D}$, reference model $f_{\text{ref}}$
Output: $f_{\text{int}} = \mathbf{W} \circ \Phi$
Initialize environment matrix $\mathbf{M}_{n_1}$, $\mathbf{M}_{n_2}$
Update $\mathbf{M}_{n_1}$, $\mathbf{M}_{n_2}$ with Eq. 5 and Eq. 6
Initialize $f_{\text{ref}}$
for $X$ in $\mathcal{D}$ do
    Get environment $e \in \mathcal{E}_{n_1}$ of $X$ from $\mathbf{M}_{n_1}$
    Update $\Phi$ and $\mathbf{W}_e$ with Eq. 7
end for
for $X$ in $\mathcal{D}$ do
    Get environment $e \in \mathcal{E}_{n_2}$ of $X$ from $\mathbf{M}_{n_2}$
    Update $\mathbf{W}_e$ with Eq. 8
end for
```

4 Experiment

4.1 NLU Tasks and Benchmarks

We apply our method on three NLU tasks to evaluate the effectiveness of our method. Specifically, we train on the original training set and evaluate on both the IID and the OOD evaluation sets. The accuracy is reported for all the benchmark datasets.

Natural Language Inference aims to classify the relationship between two sentences, *i.e.*, a premise and a hypothesis, into three classes: “entailment”, “contradiction” and “neutral”. It has been observed that NLI models may rely on the lexical overlap bias (McCoy et al., 2019). We adopt

Method	MNLI		FEVER		QQP	
	IID	**OOD**	**IID**	**OOD**	**IID**	**OOD**
	Dev	***HANS***	**Dev**	***Symmetric***	**Dev**	***PAWS***
Naïve Fine-tuning	84.5	62.4	85.6	63.1	91.0	33.5
Reweighting (KB)	83.5	69.2	84.6	66.5	89.5	50.8
Product-of-Expert (KB)	82.9	67.9	86.5	66.2	88.8	58.1
Learned-Mixin	84.0	64.9	83.1	64.9	86.6	56.8
Regularized-Confidence (KB)	84.5	69.1	86.4	66.2	89.0	36.0
Reweighting (UB)	82.3	69.7	87.1	65.5	85.2	57.4
Product-of-Expert (UB)	81.9	66.8	85.9	65.8	86.1	56.3
Regularized-Confidence (UB)	84.3	67.1	87.6	66.0	89.0	43.0
Forgettable Examples	83.1	70.5	87.1	67.0	89.0	48.8
Self-Debiasing	83.2	71.2	-	-	90.2	46.5
EIIL	83.9	69.9	89.2	68.1	87.9	57.3
BAI (Ours)	82.3$_{\pm 0.7}$	**72.7**$_{\pm 0.9}$	90.1$_{\pm 0.5}$	**69.1**$_{\pm 0.4}$	84.2$_{\pm 1.2}$	**65.0**$_{\pm 1.7}$

Table 1: Comparing our method to SOTAs on three benchmarks. Performance shown is in terms of accuracy. "KB" and "UB" denote known bias version and unknown bias version respectively. Results of Naive Fine-tuning, Reweighting, Product-of-Expert, Learned-Mixin and Regularized-Confidenceand with known bias are from Ghaddar et al. (2021), Utama et al. (2020b) and Utama et al. (2020a). Results of others are from the original paper (see Section 4.3).

MNLI (Williams et al., 2018) and HANS (McCoy et al., 2019) as the IID and OOD sets, respectively.

Fact Verification also takes in a pair of sentences, *i.e.*, a claim and an evidence, and requires the model to give the position of the evidence towards the claim. The labels are "support", "refutes", and "not enough information". Fact verification models often suffer from the claim-only bias (Utama et al., 2020b). In this paper, we use FEVER (Thorne et al., 2018) as the IID data and FEVER Symmetric (Schuster et al., 2019) as the OOD data.

Paraphrase Identification identifies whether a sentence is paraphrase of another sentence. A sentence pair is labeled as "duplicate" if the two sentences share the same semantic meaning, otherwise "non-duplicate". Similar to NLI, lexical overlap bias exists in paraphrase identification. We use QQP (Wang et al., 2018) in training as the IID set and PAWS (Zhang et al., 2019) as the OOD set.

4.2 Implementation

BERT-base (Devlin et al., 2019) from Hugging-Face's Transformers (Wolf et al., 2020) is deployed as the feature extractor for fair and direct comparison with previous methods. The reference model is also based on BERT-base which is the same as in Devlin et al. (2019), *i.e.*, one classifier layer on top of BERT. For standard hyperparameters for the training of NLU model, we use the same configuration as Utama et al. (2020a,b), *i.e.*, 3 epochs of training, learning rate of $5e-5$ for NLI and $2e-5$ for fact verification and paraphrase identification. Unlike previous methods (Clark et al., 2019; Grand and Belinkov, 2019; Clark et al., 2020; Sanh et al., 2020; Ghaddar et al., 2021) which are directly evaluated on the OOD set, we only perform checkpoint selection on the OOD set. We choose hyperparameters exclusive to our method according to the analysis on the NLI task (see RQ3) and deploy the same configuration for the other two tasks to avoid hyperparameter tuning. Specifically, we set the learning rate to $1e-2$ for automatic stratification to optimize the environment matrix, and $n_1 = 5$ and $n_2 = 2$ for bottom-up intervention. We also fix λ to $1e2$. Note the coarse-grained partition may require multiple turns of training to achieve better performance. The average results over 5 runs with different random seeds are reported.

4.3 Comparison with SOTAs

In this section, we compare our method with the following baselines: **Naïve Fine-tuning** (Devlin et al., 2019) directly fine-tunes the pre-trained language model on the downstream NLU tasks; **Reweighting** (Clark et al., 2019) reweights each training sample according to the confidence on bias model; **Product-of-Expert** (Hinton, 2002) trains the robust model fused with the bias model by sum of logits; **Learned-Mixin** (Clark et al., 2019) utilizes a different fusion method. **Regularized-Confidence** (Utama et al., 2020a) enhances the model in a knowledge distillation fashion; Unknown bias version methods in Utama et al. (2020b) adopt the bias model trained only with a small number of data; **Forgettable Examples** (Yaghoobzadeh et al., 2021) trains the model with an additional round with the forgotten data; **Self-Debiasing** (Ghaddar et al., 2021) utilizes bottom layers of model as the bias model; **EIIL** (Creager et al., 2021) is the IRM method that inspired this paper, which is originally applied to CV.

Table 1 summarizes the performance comparison between BAI and the above SOTA methods. Overall, BAI achieves the top performance on all the OOD sets. Specifically, BAI significantly outperforms naïve Fine-tuning by doubling the accuracy on PAWS (65.0% vs. 33.5%), which demonstrates that BAI with causality-theoretic basis is effective for OOD generalization on NLU tasks. Also, BAI surpasses SOTA methods with 6.9% gains over previous best result on PAWS, which shows the superiority of BAI over reweighting based methods.

We also observe a trade-off between IID and OOD on MNLI and QQP across most of the methods, i.e., performance gains on OOD are achieved with the sacrifice of IID performance. It is because naïve fine-tuning fits IID training data well. Interestingly, the IID test data of FEVER benefits from debiasing methods, which suggests that the data distribution of the IID test data may be different from that of the training data.

4.4 Ablation Studies

In this section, we conduct extensive ablation studies to evaluate the components in our BAI and answer the following research questions.

RQ1: *How does each component of BAI contribute to the performance gains?*

Answer: We design four ablative settings: (a) Replacing the learned environment matrix with a randomly initialized one; (b) Removing the regularizer term in Eq. 7 and 8; (c) Replacing bottom-up intervention with single intervention, *i.e.*, removing Eq. 8. (d) Using the same number of classifiers on naïve fine-tuning model as our BAI.

As reported in Table 2, the settings (a) and (d) prove that the environment partition is vital in our method and the improvement of our method is not from the added parameters[5]. Result of (b) reveals that both the regularizer term and the design of one classifier for one environment contribute to the gains in our method. Finally, the full method with bottom-up intervention outperforms (c), which demonstrates the effectiveness of multi-granular intervention.

Ablative Setting	**Dev**	***HANS***
Naïve FT	84.5	62.4
(a) Randomized Environment	84.0	62.4
(b) w/o Regularizer	83.0	66.8
(c) One Intervention	83.9	69.9
(d) Naive FT+Multiple Classifiers	84.4	62.6
Full Method	82.3	**72.7**

Table 2: **RQ1.** Results of ablative settings on MNLI. "FT" denotes Fine-tuning.

Stratifying Method	**Dev**	***HANS***
No Stratifying	84.5	62.4
(1) Domain Information	84.2	63.2
(2) Confidence	84.0	67.7
(3) Lexical Overlap	83.8	65.6
Automatic Stratifying (Ours)	83.9	**69.9**

Table 3: **RQ2.** Results of alternative methods for environment stratification on MNLI.

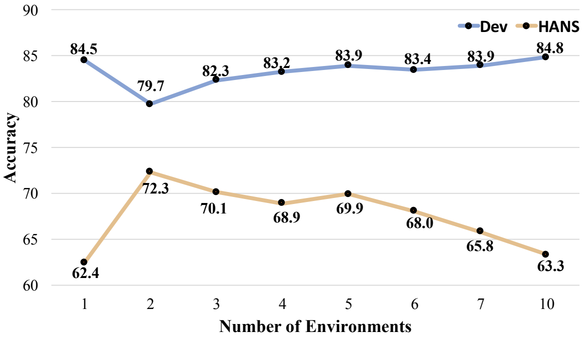
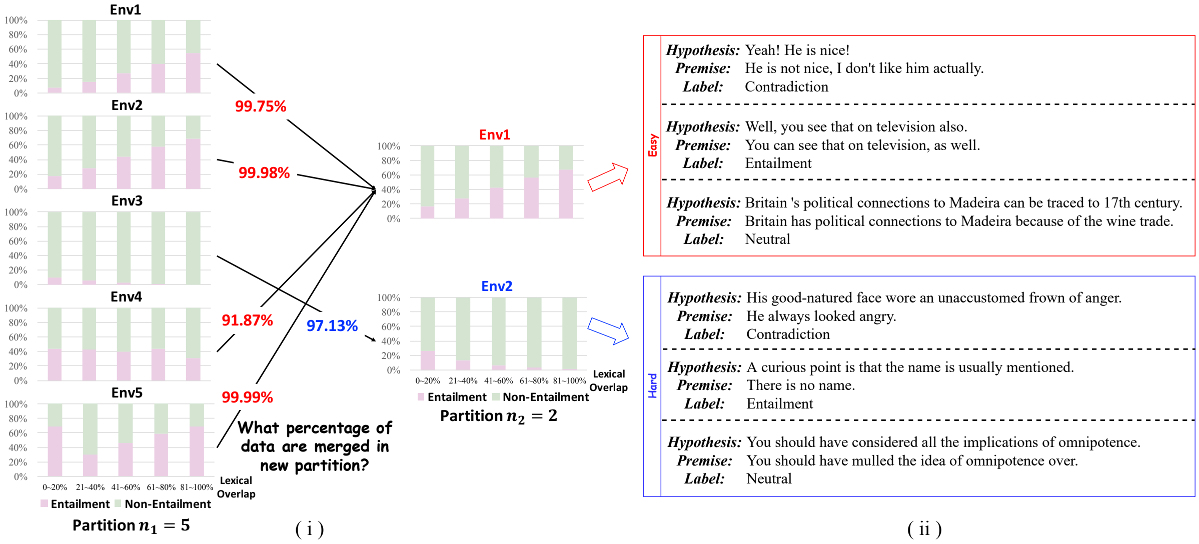


Figure 4: **RQ3.** The accuracies of one round of intervention on MNLI with different numbers of environments.

RQ2: *Is there any other solution for stratification?*

Answer: Yes. We evaluate several alternative methods for partition on MNLI according to the attached information of training samples: (1) Domain information, *i.e.*, "fiction", "governmnet", "slate", "telephone" and "travel"; (2) Confidence of predic-

[5] BAI introduces 0.008% more parameters compared to that of Naïve Fine-tuning.



Hypothesis: Yeah! He is nice!
Premise: He is not nice, I don't like him actually.
Label: Contradiction

Hypothesis: Well, you see that on television also.
Premise: You can see that on television, as well.
Label: Entailment

Hypothesis: Britain 's political connections to Madeira can be traced to 17th century.
Premise: Britain has political connections to Madeira because of the wine trade.
Label: Neutral

Hypothesis: His good-natured face wore an unaccustomed frown of anger.
Premise: He always looked angry.
Label: Contradiction

Hypothesis: A curious point is that the name is usually mentioned.
Premise: There is no name.
Label: Entailment

Hypothesis: You should have considered all the implications of omnipotence.
Premise: You should have mulled the idea of omnipotence over.
Label: Neutral

Figure 5: **RQ4.** (i) Characteristics and relationship for two partitions. Each sub-graph shows the same analysis setting as in Figure 1 in corresponding environment; (ii) Examples for the easy and hard samples for the partition with $n_2 = 2$.

Order & Combination	**Dev**	***HANS***
$\mathcal{E}_2 \rightarrow \mathcal{E}_5$	81.7	70.1
$\mathcal{E}_5 \rightarrow \mathcal{E}_3$	83.7	71.4
$\mathcal{E}_5 \rightarrow \mathcal{E}_3 \rightarrow \mathcal{E}_2$	81.3	73.5
$\mathcal{E}_5 \rightarrow \mathcal{E}_2$ (Config in Table 1)	81.1	73.3

Table 4: **RQ3.** Results of different orders and combinations of environment numbers on MNLI, arrows represent the intervention order.

tion (Clark et al., 2019). We calculate the highest confidence or the options and the confidence for the ground-truth label. All the samples are grouped into environments by K-Means (Hartigan and Wong, 1979) according to the two confidence scores; (3) Prior knowledge of bias, *i.e.*, lexical overlap bias in Figure 1. We also group them into different environments by K-Means. For fairness, we fix the number of environments as 5, which is the number of domains in the setting (1). We compare the above settings with our model trained using only one intervention in Eq. 7.

As summarized in Table 3, the results show that directly using domain information as basis for environments stratifying has very few gains, *i.e.*, 0.8%. Although intervention based on domain information is beneficial for every domain, such intervention does not provide a good partition for debiasing as the lexical bias still exists. Stratifying based on confidence and lexical overlap shows considerable improvements compared to that of no stratifying, which demonstrates the two factors are indeed related to the confounder of MNLI. Note that the automatic stratifying method is designed for unobserved confounder, which outperforms the simple heuristics in settings (2) and (3) without using any prior knowledge of bias.

RQ3: *How to set the number of environments?*
Answer: We first analyze the situation of only one round of intervention and visualize the performance trend in Figure 4. Note that setting the number of environments as one equals to naïve fine-tuning, *i.e.*, no stratification. Overall, there is a trade-off in the results between Dev and HANS, *i.e.*, IID and OOD performances. This phenomenon is particularly prominent in $\mathcal{E}_2$. The reason is that only one intervention forces the model to focus on only one confounder. In this case, it forces the model to pay much attention on the harder samples, *i.e.*, the confounder of crowdsourced worker preference, leading to significant performance drop on dev set (see RQ4 for more details).With the number of environments increasing, the gaps between the environments are also smaller, *i.e.*, the OOD performance of ten environments is close to that of the naïve fine-tuning.

We further analyze the multiple interventions. We conduct experiments with the number of interventions in different orders or combinations. The experiment results are summarized in Table 4. We observe that applying the partition with two environments in the final intervention is better and increasing the turns of intervention only brings

Method	MNLI		FEVER		QQP	
	IID	OOD	IID	OOD	IID	OOD
Naïve F.T.	87.3	69.8	87.1	68.6	90.8	38.4
BAI	85.3	**76.7**	91.2	**72.9**	83.5	**70.2**

Table 5: **RQ5.** Results of BAI with RoBERTa as backbone. "Naïve F.T." denotes Naïve Fine-tuning

marginal improvements. Thus, we simply fix $\mathcal{E}_5 \rightarrow \mathcal{E}_2$ for all tasks in our paper.

RQ4: *What is each environment like?*

Answer: Figure 5 inspects each environment in two partitions, *i.e.*, $\mathcal{E}_5$ and $\mathcal{E}_2$, on MNLI and summarizes the characteristic for each environment. $\mathcal{E}_2$ can be regarded as a coarse variant of $\mathcal{E}_5$, *i.e.*, the first environment of $\mathcal{E}_2$ partition combines four environments of $\mathcal{E}_5$. We can see that both partitions contain environments with distinct characteristics. $\mathcal{E}_2$ focuses more on crowdsourced worker preference while $\mathcal{E}_5$ shows each environment with more diverse situation for the nature bias, *i.e.*, lexical overlap bias.

We further investigate the crowdsourced worker preference in $\mathcal{E}_2$, *i.e.*, the difficulty of the samples in these two environments is distinguishable. Samples in the second environment are more challenging compared to the first one. As depicted in Figure 5 (ii), reasoning of easy samples is straightforward, *i.e.*, `nice` versus `not nice` and `do not like`. In contrast, hard examples require a deep understanding of the semantic meaning. For instance, the hard samples with contradiction and entailment as labels expect the model to have the ability to identify the current situation, *e.g.*, `no name` for now, and the usual situation, *e.g.*, `name is usually mentioned` in the past. The above inspection reveals that BAI helps to generate meaningful and multifactorial partition.

RQ5: *Whether BAI is model-agnostic?*

Answer: We apply the same hyperparameters and partitions on BERT to RoBERTa (Liu et al., 2019). The results in Table 5 demonstrate that the proposed BAI can be applied on more advanced language model than BERT.

5 Conclusions

In this paper, we explore how to improve the robustness of NLU models under OOD setting, and propose a bottom-up automatic intervention method for debiasing. The experiment results demonstrate the superiority of our model over state-of-the-art methods. In future work, we will consider two improvements on BAI. First, we target at an end-to-end framework for intervention and dynamic learn the partition of environment for NLU tasks. Second, we want to ease the trade-off effect between IID and OOD sets.

6 Limitations

The limitations of this paper are twofold. First, the proposed method is only evaluated on natural language understanding tasks. Thus the effectiveness on natural language generation tasks and sequence labeling tasks is not guaranteed. Similarly, the optimal hyper-parameters for other tasks may also differ from the selections stated in this paper. Second, the performance trade-off (see Table 1) is non-negligible on the IID set compared to the OOD set. It is not desirable when the model is applied to the normal scenario, *e.g.*, the confounders provide shortcuts for model inference.

Acknowledgments

The authors gratefully acknowledge the support by the Lee Kuan Yew Fellowship awarded by Singapore Management University.

References

Hossein Amirkhani and Mohammad Taher Pilehvar. 2021. Don't discard all the biased instances: Investigating a core assumption in dataset bias mitigation techniques. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4720–4728.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32:841–852.

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

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3031–3045.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2021. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

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

Gabriel Grand and Yonatan Belinkov. 2019. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *NAACL HLT 2019*, page 1.

Judith A Hall, Michael A Milburn, and Arnold M Epstein. 1993. A causal model of health status and satisfaction with medical care. *Medical care*, pages 84–94.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142.

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. 2021. Heterogeneous risk minimization. *arXiv preprint arXiv:2105.03818*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695.

Yulei Niu, Long Chen, Chang Zhou, and Hanwang Zhang. 2022. Respecting transfer gap in knowledge distillation. *Advances in Neural Information Processing Systems*.

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

Yulei Niu and Hanwang Zhang. 2021. Introspective distillation for robust question answering. *Advances in Neural Information Processing Systems*, 34:16292–16304.

Judea Pearl. 1993. [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Judea Pearl. 2010. Causal inference. *Causality: Objectives and Assessment*, pages 39–58.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Hachette UK.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10860–10869.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. 2020. The risks of invariant risk minimization. In *International Conference on Learning Representations*.

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

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725.

Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1417–1427.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610.

V Vapnik. 1991. Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, pages 831–838.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, Timothy J Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2021. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sicheng Yu, Yulei Niu, Shuohang Wang, Jing Jiang, and Qianru Sun. 2020. Counterfactual variable control for robust and interpretable question answering. *arXiv preprint arXiv:2010.05581*.

Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional few-shot learning. *Advances in Neural Information Processing Systems*, 33.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021. De-biasing distantly supervised named entity recognition via causal intervention. *arXiv preprint arXiv:2106.09233*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. 2022. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3589–3597.