SMU Singapore Management University
School of Computing and Information Systems

達摩院 ALIBABA DAMO ACADEMY

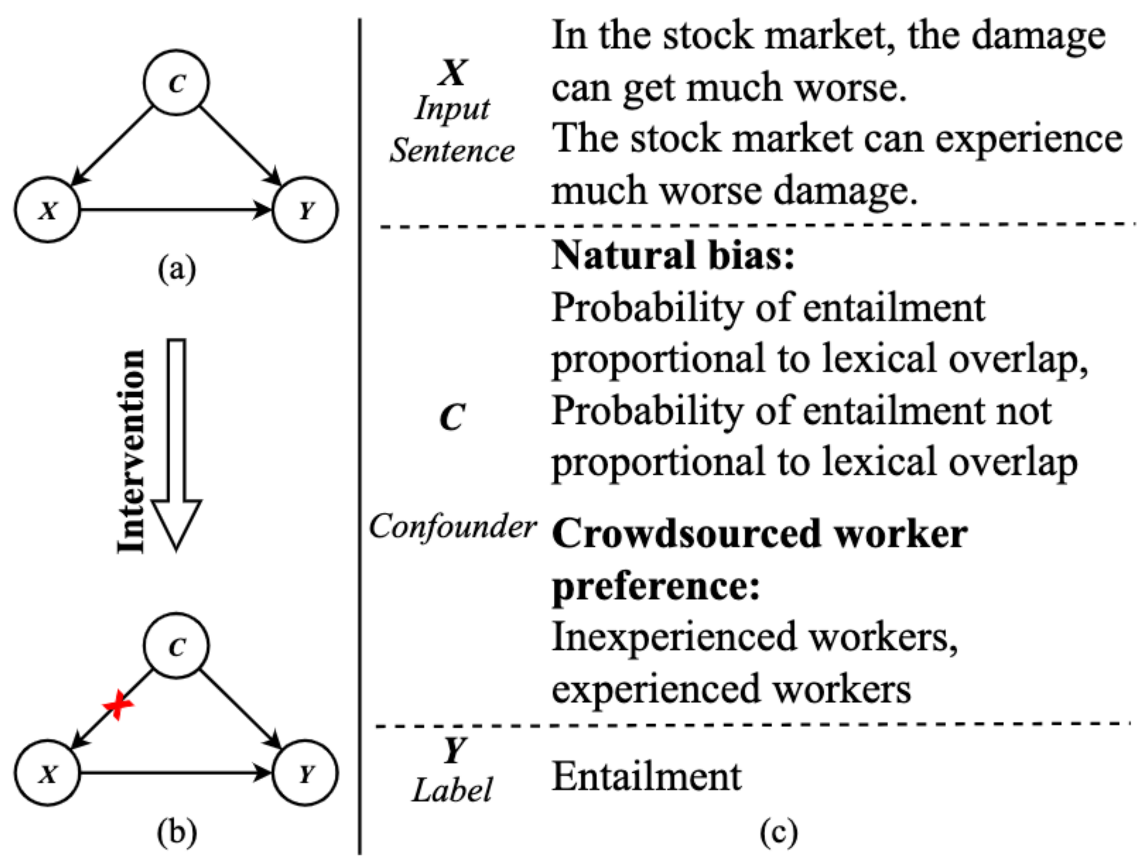
Interventional Training for Out-Of-Distribution Natural Language Understanding

Sicheng Yu, Jing Jiang, Hao Zhang, Yulei Niu, Qianru Sun, Lidong Bing

Motivation

- Natural Language Understanding (NLU) models often suffer in Out-Of-Distribution (OOD) settings.
- Most of previous debiasing methods rely on the known bias and sample reweighting.
- What we target: a debiasing method without sample reweighting for unknown bias.

Our Approach



- Taking NLI as example, we analyze the vulnerability of model from the view of causality.
- The unveiled crux is *confounding bias* and a common solution for de-confounding is *intervention* with two implementing challenges: the confounder C is unobserved multifactorial.
- For the first challenge (unobserved confounder), we propose to automatically stratify the data into environments by maximizing the difference of data across the environments.
- For the second challenge (multifactorial confounder), we propose bottom-up intervention for multi-granular de-confounding.

Experiment Results

Method	MNLI		FEVER		QQP	
	IID Dev	OOD *HANS*	IID Dev	OOD *Symmetric*	IID Dev	OOD *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)	$82.3_{\pm 0.7}$	$\mathbf{72.7}_{\pm 0.9}$	$90.1_{\pm 0.5}$	$\mathbf{69.1}_{\pm 0.4}$	$84.2_{\pm 1.2}$	$\mathbf{65.0}_{\pm 1.7}$

Our method named Bottom-up Automatic Intervention (BAI) outperforms the SOTA debiasing methods based on bias model and sample reweighting on three different tasks and OOD settings..

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**

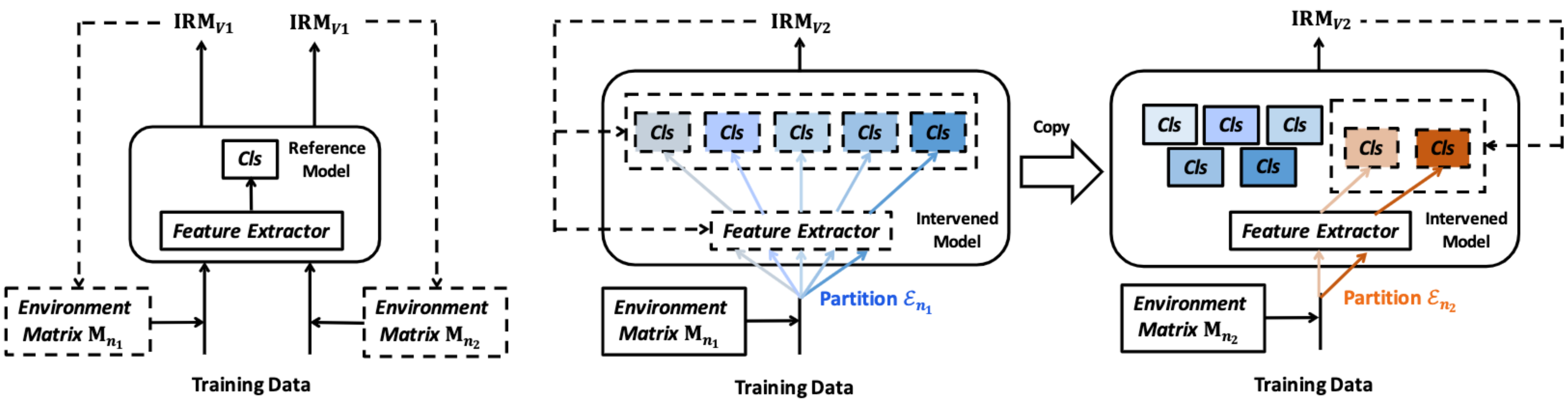
Table 3: **RQ2.** Results of alternative methods for environment stratification on MNLI.

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.

- The ablative studies on different stratifying methods (left figure) demonstrate that the proposed automatic stratification is superior to rule-based alternatives.
- The ablative studies on different orders of partitions show that the bottom-up order for intervention is better than other orders.

Solution Architecture



(a) Automatic Stratifying (b) Bottom-up Intervention

Conclusions

- We explore how to improve the robustness of NLU models under OOD setting, and propose a bottom-up automatic intervention method.
- The experiment results demonstrate the superiority of our method over state-of-the-art methods on three benchmarks.