Counterfactual Adversarial Learning with Representation Interpolation

Wei Wang, Boxin Wang, **Ning Shi**, Jinfeng Li, Bingyu Zhu, Xiangyu Liu, Rong Zhang

{luyang.ww, **shining.shi**, jinfengli.ljf,zhubingyu.zby, eason.lxy, stone.zhangr}@alibaba-inc.com, boxinw2@illinois.edu

Introduction

- Deep learning models exhibit a preference for statistical fitting over logical reasoning, which severely limits the model performance, especially in small data scenarios.
- We propose CAT, an end-to-end and task-agnostic Counterfactual Adversarial Training framework to tackle the problem using causal inference.

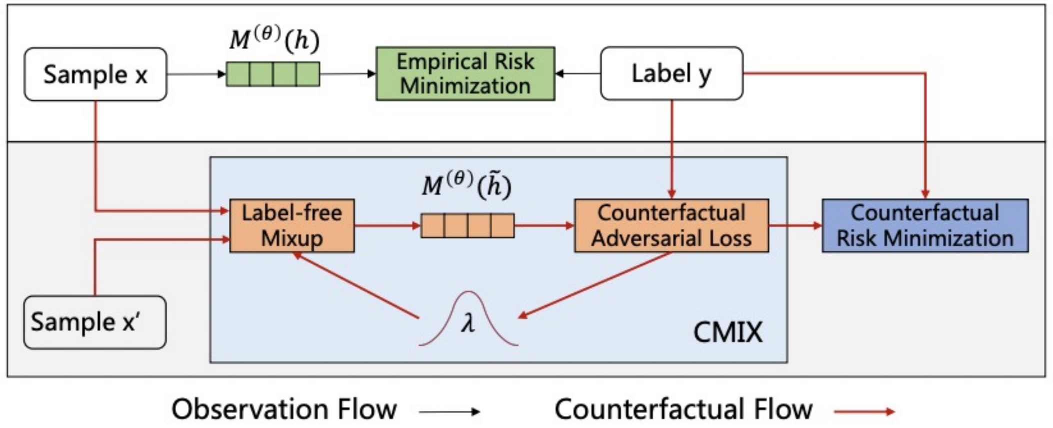


Figure 1: The framework of CAT. Besides the normal supervised ERM (Observation) flow on the top, for a certain observation x, CAT will randomly sample another x' from training data. Then a counterfactual representation $\tilde{h}$ is generated and optimized by CMIX. Finally, CRM is applied on final model output $M^{(\theta)}(\tilde{h})$.

Contributions

- We investigate the problem of spurious correlations from a causality perspective which has not been widely studied in conventional statistical learning.
- We propose CMIX for counterfactual representation interpolation to approximate do-calculus realization in a deep learning framework, which is adaptively optimized by a novel Counterfactual Adversarial Loss.
- We show that CAT outperforms SOTA by a large margin across different tasks particularly when data is limited.

Methods

- label-free mixup: conducts do-calculus and generates counterfactual representations by interpolating the hidden states to generate counterfactual representaitons.
- We propose Counterfactual Adversarial Loss (CAL) to further optimize the counterfactual representations.
- CRM is designed to enable the model to learn from both original representations and counterfactual ones.

Model	Yahoo! Answers				IMDB				SNLI			
	10	50	250	1000	10	50	250	1000	10	50	250	1000
$\text{BERT}_{\text{BASE}}$	61.02	66.39	70.07	72.33	73.28	78.03	82.38	85.88	42.68	57.62	70.17	77.16
TMix	62.19	67.01	70.15	72.30	74.32	78.64	82.58	85.90	43.90	58.55	70.57	77.40
CAT *	62.34	67.20	70.11	72.29	73.77	78.98	82.45	85.96	44.37	59.42	71.23	77.89
CAT	**63.53**	**68.11**	**71.40**	**72.52**	**75.55**	**80.13**	**83.15**	**86.11**	**46.23**	**60.27**	**72.13**	**78.20**
$\text{RoBERTa}_{\text{BASE}}$	61.95	66.96	69.61	71.21	81.57	84.30	87.00	88.36	40.72	59.92	77.96	83.09
CAT *	63.09	**67.84**	70.08	71.95	82.80	85.11	87.40	88.45	**41.95**	63.33	79.15	83.25
CAT	**63.55**	67.78	**70.45**	**72.02**	**83.25**	**85.12**	**87.50**	**88.93**	41.30	**64.47**	**79.69**	**83.75**
$\text{BERT}_{\text{LARGE}}$	63.54	67.96	70.75	72.93	76.51	81.22	85.42	**87.32**	**44.33**	60.10	74.02	81.04
CAT *	64.33	68.07	70.72	72.95	**76.97**	81.05	85.38	86.93	43.07	62.80	75.97	81.18
CAT	**64.73**	**68.15**	**70.95**	**73.06**	75.10	**82.52**	**86.02**	87.00	43.83	**64.77**	**76.77**	**81.67**
$\text{RoBERTa}_{\text{LARGE}}$	64.38	67.80	70.60	72.28	81.50	87.63	89.03	90.06	38.22	62.73	82.27	85.99
CAT *	66.20	68.92	71.10	72.90	79.95	87.55	89.48	90.10	39.15	61.85	82.90	85.63
CAT	**66.30**	**69.28**	**71.25**	**73.30**	**84.80**	**88.55**	**89.85**	**90.10**	**40.33**	**65.07**	**83.15**	**86.05**

Table 1: The average accuracy after multiple runs on Yahoo! Answers, IMDB and SNLI datasets. Bellowing the individual dataset is the number of training samples per class.

Model	SQuAD 1.1			SQuAD 2.0		
	1/20	1/10	1/5	1/20	1/10	1/5
$\text{BERT}_{\text{BASE}}$	51.83/62.50	66.06/76.56	72.25/81.75	51.10/54.12	55.60/58.84	61.84/65.42
CAT *	**63.90/74.93**	69.36/79.44	74.10/83.34	55.44/57.55	**59.84/62.44**	61.77/64.97
CAT	62.71/74.14	**69.49/79.44**	**74.33/83.43**	**56.22/58.47**	59.71/62.44	**63.26/66.72**
$\text{BERT}_{\text{LARGE}}$	70.66/81.29	75.85/85.16	79.14/87.24	59.41/63.03	66.28/70.30	**71.30/74.88**
CAT *	72.18/82.15	75.69/84.83	79.06/87.08	61.84/65.27	66.55/70.08	69.40/72.87
CAT	**72.30/82.17**	**76.37/85.09**	**79.18/87.28**	**61.82/65.32**	**67.38/70.79**	69.31/72.37

Table 2: The model performance of EM/F1 on SQuAD 1.1 and SQuAD 2.0. Bellowing the individual dataset is the proportion of full training data used.

Source Code

https://github.com/ShiningLab/CAT

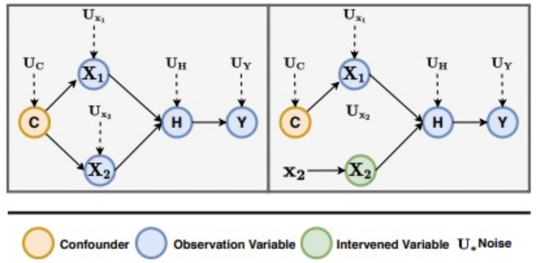
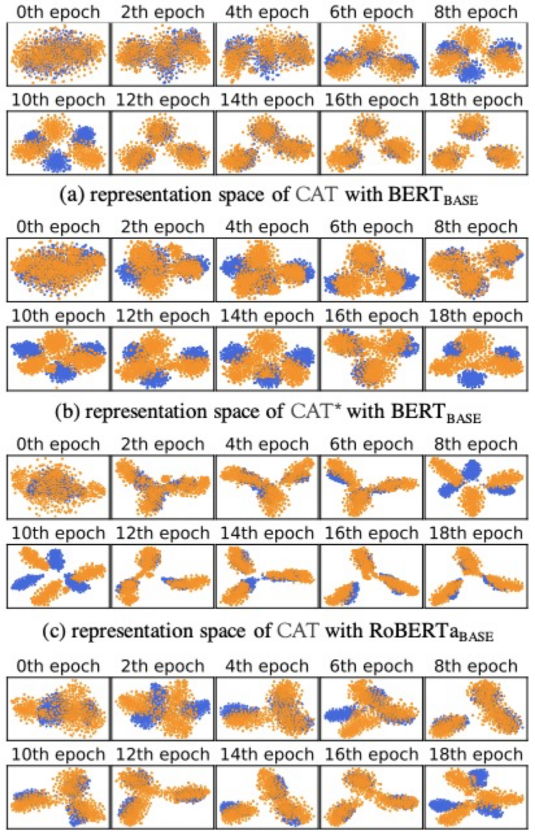


Figure 2: SCM of data generation mechanism. Left: Spurious correlations exist between $\mathbf{X}_1$ and $\mathbf{X}_2$ in observation data caused by confounder $\mathbf{C}$. Right: Confounder is eliminated by *do-calculus*.



(a) representation space of CAT with $\text{BERT}_{\text{BASE}}$

(b) representation space of CAT* with $\text{BERT}_{\text{BASE}}$

(c) representation space of CAT with $\text{RoBERTa}_{\text{BASE}}$

(d) representation space of CAT* with $\text{RoBERTa}_{\text{BASE}}$

Figure 4: Representation space visualization through tSNE for CAT and CAT *. during the training process on SNLI data with 250 samples per class. (a) and (b) represent CAT and CAT * on $\text{BERT}_{\text{BASE}}$ and (c) and (d) for $\text{RoBERTa}_{\text{BASE}}$