Counterfactual Adversarial Learning with Representation Interpolation

Wei Wang[1] Boxin Wang[2] Ning Shi[1]

Jingfeng Li[1] Bingyu Zhu[1] Xiangyu Liu[1] Rong Zhang[1]

{luyang.ww, shining.shi, jinfengli.ljf,

zhubingyu.zby, eason.lxy, stone.zhangr}@alibaba-inc.com

boxinw2@illinois.edu

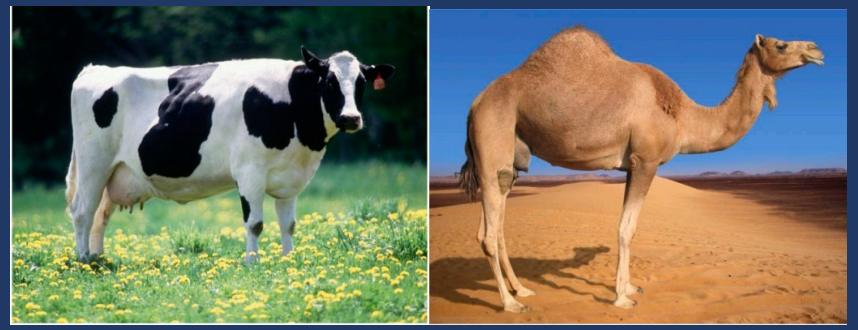
[1]Alibaba Group [2]University of Illinois at Urbana-Champaign

Motivation

Conventional Machine Learning is more like statistical fitting rather than logical reasoning.

Camels or Cows?

Train set



Test Set

Statistical Correlation:

Cows ⟷ Grassy background
Camels ⟷ Sandy background

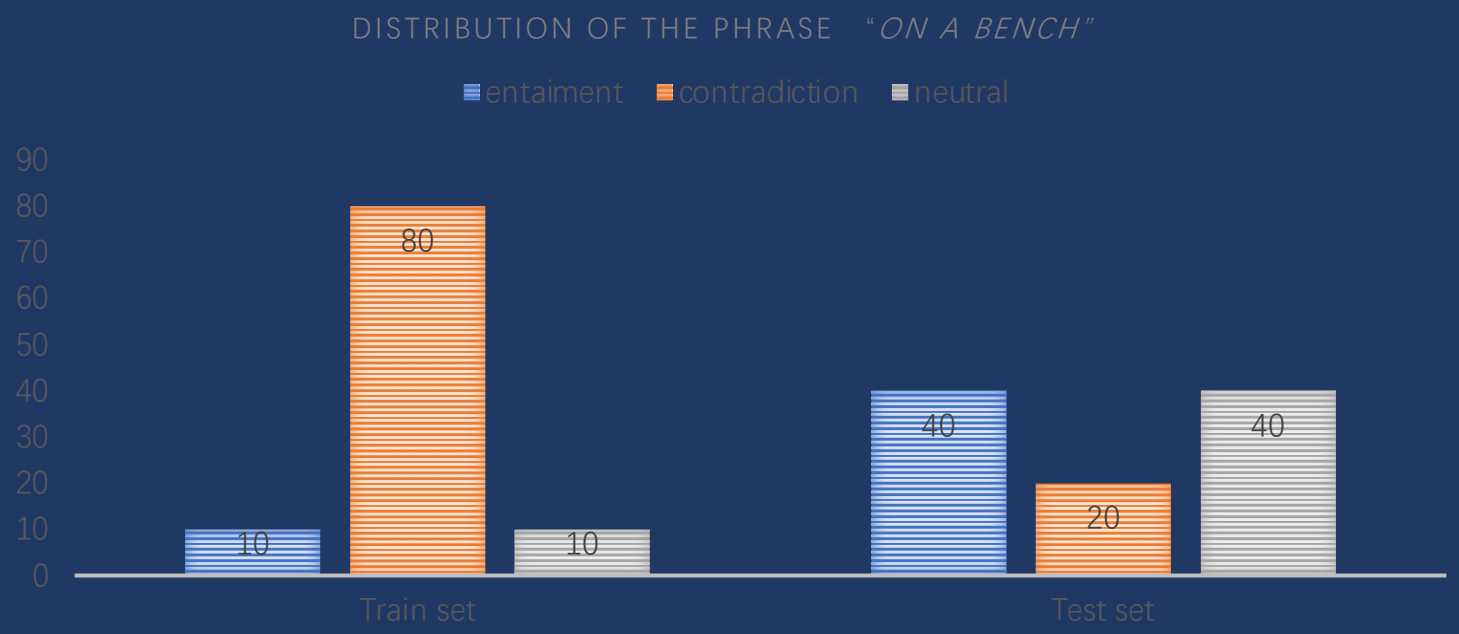
Out of distribution Test data:

Cows ⤫ Grassy background
Camels ⤫ Sandy background

Statistical Correlations will become spurious bias on OOD data thus undermine model performance.

Motivation

Such spurious bias is also common in NLP tasks.



- Such skewed distribution in training data indicates a statistical correlation between *on a bench* and label *contradiction*.
- The reason can be attributed to some common cause (Confounder) of both labels and the phrase, such as subjective bias of human annotator, the domain of data, the region where data is collected, etc.
- Models tend to use such correlations as shortcuts to help predicting labels.
- *on a bench* do not lead to certain labels in NLI task logically, thus such correlation is non causal and may change in test set.
- Models will fail on the test set where such bias disappear or change.

Counterfactual Adversarial Training (CAT)

Target:

Find the minimal intervention that will alter the model predictions and let the model to learn from such representations. Such intervention will alleviate the spurious bias thus let model explore the causal effect rather than simple correlations.

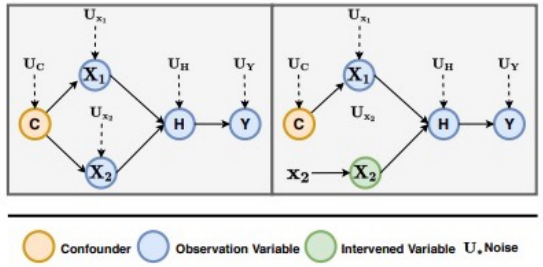


Figure 2: SCM of data generation mechanism. Left: Spurious correlations exist between $\mathbf{X}_1$ and $\mathbf{X}_2$ in observation data caused by confounder $\mathbf{C}$. Right: Confounder is eliminated by *do-calculus*.

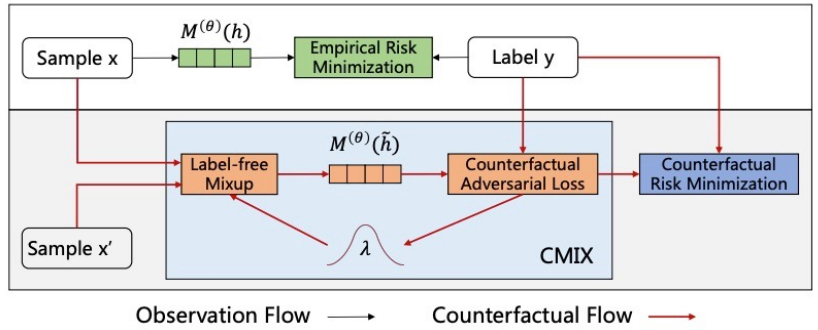


Figure 1: The framework of CAT. Besides the normal supervised ERM (Observation) flow on the top, for a certain observation x, CAT will randomly sample another x' from training data. Then a counterfactual representation $\tilde{h}$ is generated and optimized by CMIX. Finally, CRM is applied on final model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

To alleviate the influence of confounders, we interpolate in hidden space of transformers as interventions to generate virtual representations. Next a counterfactual adversarial loss (CAL) aligned with the definition of counterfactuals are optimized to find the balanced trade-off between intervention and model prediction changes. Finally, we apply Counterfactual Risk Minimization Principle to enable model to learn from generated representations.

Counterfactual Adversarial Training (CAT)

Algorithm Details

Intervention: a label-free interpolation designed for transformers and interpolates in hidden space to avoid the discrete nature of textual data.

$$\tilde{h}_m^{(i)} = \lambda^{(i)} h_m^{(j)} + (1-\lambda^{(i)}) h_m^{(i)}$$
$$\tilde{h}_l^{(i)} = M_l^{(\theta)}(\tilde{h}_{l-1}^{(i)}), \quad l \in \{m+1, ..., L\}, \tag{4}$$

CAL: an adversarial trade-off game for minimizing intervention and maximizing label change.

$$\arg\max_{\lambda^{(i)}} -\left\|\lambda^{(i)}\right\|_p + \gamma L(M^{(\theta)}(\tilde{h}^{(i)}), y^{(i)}) + \eta \Phi(M^{(\theta)}(\tilde{h}^{(i)})), \tag{5}$$

CRM: a learning principle that allows model to learn from both original representations and counterfactual ones.

$$\hat{R}_c(M^{(\theta)}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\Phi(M^{(\theta)}(h^{(i)}))}{\Phi(M^{(\theta)}(\tilde{h}^{(i)}))} L(M^{(\theta)}(h^{(i)}), y^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\omega}(h^{(i)}) L(M^{(\theta)}(h^{(i)}), y^{(i)}). \tag{9}$$

Algorithm 1: Counterfactual Adversarial Training Approach (CAT)

Input: Dataset $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, model $M^{(\theta)}$, mixup layer candidate set $\mathcal{Q}$, Beta distribution parameters α and β, denote as $Beta(\alpha, \beta)$, couterfactual adversarial loss iteration step $\mathcal{L}$, warm up step $\mathcal{K}$, max step $\mathcal{T}$

for *step* $k \in \{0, 1, ..., \mathcal{K}\}$ **do**
- Sample one batch $X^{(k)} \in D$. Denote corresponding representations as $h^{(k)}$; Do ERM on $M^{(\theta)}(h^{(k)})$;

for *step* $t \in \{0, 1, ..., \mathcal{T}\}$ **do**
- Sample one batch $X^{(t)}$. Denote corresponding representations as $h^{(t)}$;
- For each $x^{(i)}$ in $X^{(t)}$, random sample $q \in \mathcal{Q}$ and $\lambda^{(i)} \sim Beta(\alpha, \beta)$ and generate mixed representations in latent space using (Eq.4) to get one batch of counterfactual representations $\tilde{h}^{(t)}$;
- **for** $l \in \{0, 1, ...\mathcal{L}\}$ **do**
  - Optimize counterfactual representations using CAL (Eq.5);
- Do CRM on $M^{(\theta)}(\tilde{h}^{(t)})$ and $M^{(\theta)}(h^{(t)})$;
- Do ERM on $M^{(\theta)}(h^{(t)})$;

Counterfactual Adversarial Training (CAT)

Experiments Results

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**

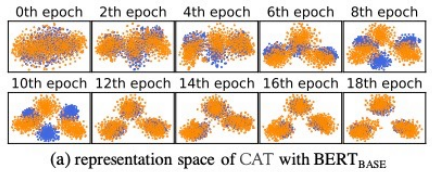
Table 1: The average accuracy after multiple runs on Yahoo! Answers, IMDB and SNLI datasets. Bellowing the individual dataset is the number of training samples per class.

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

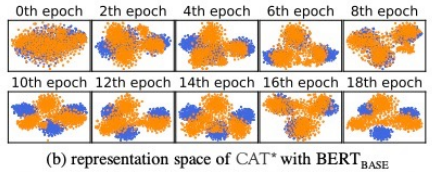
Table 2: The model performance of EM/F1 on SQuAD 1.1 and SQuAD 2.0. Bellowing the individual dataset is the proportion of full training data used.

Model	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
$\text{BERT}_{\text{BASE}}$	80.80	88.50	72.57	75.99
CAT	**81.77**	**88.98**	**74.13**	**77.36**

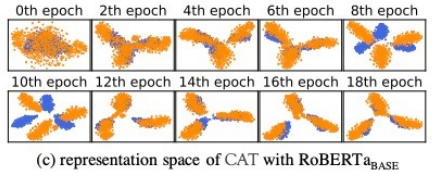
Table 3: The EM/F1 on full QA data.



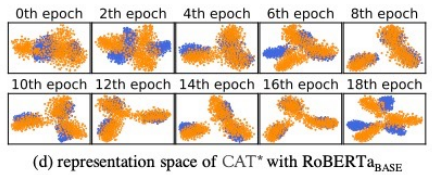
(a) representation space of CAT with $\text{BERT}_{\text{BASE}}$



(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 $\text{RoBERTa}_{\text{BASE}}$

Conclusions and Discussion

To alleviate the spurious correlation bias in training corpus and encourage causal discovery instead of simple correlations, we propose CAT from the causality perspective for introducing counterfactual representations in the training stage through latent space interpolation. Through extensive experiments on three benchmarks on the text classification, natural language inference and question answering tasks, we demonstrate that CAT is effective in promoting testing accuracy especially in the small data scenario, which outperforms SOTA baselines across different pre-trained models.

In future, we will try to extend CAT beyond pre-trained language models and transformers model by modify the intervention techniques. Also, we are working on introducing the other causal inference technique to deep learning framework to improve model robustness and stability.