# PairCFR: Enhancing Model Training on Paired Counterfactually Augmented Data through Contrastive Learning

**Xiaoqi Qiu[1]*, Yongjie Wang[2]*, Xu Guo[2], Zhiwei Zeng[2], Yue Yu[2], Yuhong Feng[1], Chunyan Miao[2]**

1. qiuxiaoqi2022@email.szu.edu.cn, yuhongf@szu.edu.cn

2. {yongjie.wang, xu.guo, zhiwei.zeng, yue.yu, ascymiao}@ntu.edu.sg

***Equal Contribution. 1. Shenzhen University, China. 2. Nanyang Technological University, Singapore.**

## 1 Introduction

- **Spurious correlations** in NLP, e.g., dataset-specific artifacts, undermine OOD generalizability
- **Counterfactually Augmented Data (CAD)** mitigates this issue by establishing direct causal relationships for models to learn more easily and more effectively.

| **ORI**ginal Example(ORI): Negative → **C**ounter**F**actual **E**xample(CFE): Positive | |
|---|---|
| ORI | After loving "Panama" (10/10), I found this one boring. ☹ |
| CFE | After loving "Panama" (10/10), I found this one funny. ☺ |

ORI ➡ Minimal edits | Label flip | Meaningful ➡ CFE

## 2 Problem

- **CAD may lead to overfitting these modifications**

gold label= Negative

The Atlantis is **supposed to be** fantastic.

prediction = Positive ✘

**How to prevent models over-relying on local edits?**

## 3 Methodology

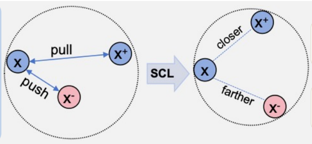
- Can the model focus more on global information?

**SCL**

**S**upervised

**C**ontrasitive

**L**earning




uniformity=> more global information

alignment=> more robust representation

**SCL effectively captures global information for alignment !**

- Enhance Model OOD generalization with Paired CAD

**PairCFR**

**Pair**wisely

**C**ounter**F**actual

Learning with

Contrastive

**R**egularization

- Pair ORI and CFE in the same batch during training
- $\mathcal{L}_{PairCFR} = \lambda \times \mathcal{L}_{CL} + (1-\lambda) \times \mathcal{L}_{std.CE}$ Synergy

$$\mathcal{L}_{CL} = -\mathbb{E}_{x_i \sim D}\mathbb{E}_{x_p \sim P_i}\left[\log \frac{e^{s_{ip}/\tau}}{e^{s_{ip}/\tau} + \sum_{x_n \in N_i} e^{s_{in}/\tau}}\right]$$

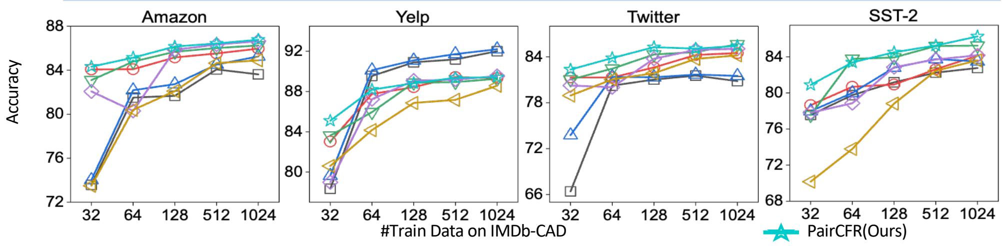
Original Negatives　　CFE (edited features)

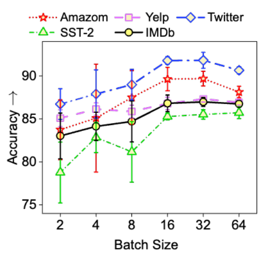
## 4 Selected Results

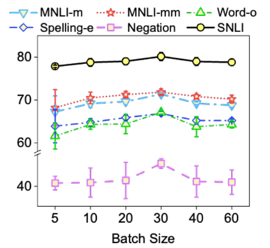
- **PairCFR** brings the highest o.o.d. performance for both SA & NLI
- CAD-based training does not always boost in-domain performance

| Method | Sentiment Analysis | | | | | | Natural Language Inference | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | In-Domain | Out-of-Dimain | | | | | In-Domain | Out-of-Dimain | | | | | |
| | IMDb | Amazon | Yelp | Twitter | SST-2 | Acc | SNLI | MNLI-m | MNLI-mm | Negation | Spelling-e | Word-o | Acc |
| RoBERTa-base | | | | | | | | | | | | | |
| ORI/Translation: Vanilla | 92.68±1.15 | 87.08±1.39 | 94.00±0.77 | 81.43±2.82 | 86.04±2.76 | 87.14 | 85.16±0.39 | 70.35±1.29 | 71.25±1.59 | 52.47±5.55 | 67.36±1.36 | 61.82±4.54 | 64.65 |
| ORI/Translation: BTSCL | **93.09±0.61** | 89.46±0.21 | **94.74±0.36** | 85.72±1.22 | 87.16±0.87 | 89.27 | **85.72±0.44** | 70.83±1.38 | 72.10±1.32 | **56.89±3.78** | 67.61±1.32 | 62.22±3.55 | 65.93 |
| ORI/Translation: CouCL | 91.22±0.83 | 89.48±0.19 | 93.04±0.58 | 87.40±0.77 | 88.07±0.66 | 89.50 | 82.37±0.52 | 70.86±1.32 | 71.38±1.23 | 51.83±2.71 | 68.08±1.23 | 64.68±1.82 | 65.37 |
| CAD: HCAD | 90.12±1.74 | 88.50±0.57 | 92.18±0.94 | 83.43±1.75 | 86.48±0.98 | 87.65 | 80.91±0.69 | 70.35±1.08 | 70.77±0.76 | 45.79±4.16 | 67.37±1.28 | 64.83±1.47 | 63.82 |
| CAD: CFGSL | 90.69±0.92 | 88.32±0.41 | 93.48±0.48 | 83.90±1.78 | 86.89±0.80 | 88.15 | 82.45±0.35 | 71.59±0.90 | 71.25±1.06 | 51.40±1.47 | 68.86±1.07 | 62.22±1.99 | 65.06 |
| CAD: ECF | 91.05±0.44 | 88.56±0.32 | 93.79±0.19 | 85.82±0.43 | 87.84±0.59 | 89.00 | 81.88±0.17 | 70.45±1.03 | 71.18±0.93 | 51.70±2.38 | 66.60±0.94 | 63.76±1.98 | 64.74 |
| CAD: PairCFR | 91.74±0.88 | **89.60±0.26** | 93.35±0.34 | **87.90±0.45** | **88.61±0.41** | **89.87** | 82.13±0.51 | **71.80±0.53** | **72.12±0.79** | 55.19±1.97 | **68.88±0.36** | **65.91±1.35** | **66.78** |

- **PairCFR** shows robustness across different data settings, especially under **few-shot settings**




- Fair #Neg ↑ => broader features
- Excessive #Neg ↑=> dilute CAD priors

## 5 Conclusion

- **PairCFR** is a simple and effective method for training models on CAD.

It demonstrates highest o.o.d. performance on both SA and NLI task.

ACL 2024 Bangkok, Thailand

SHENZHEN UNIVERSITY

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE