Causal Fine-Tuning of Pre-trained Language Models for Robust Test Time Adaptation

Jialin Yu[a,b] Yuxiang Zhou[c,d] Yulan He[c] Nevin L. Zhang[e] Junchi Yu[a] Philip Torr[a] Ricardo Silva[b]

[a] University of Oxford [b] University College London [c] King's College London [d] Queen Mary University of London [e] The Hong Kong University of Science and Technology

Motivation Example

Adapting models at test time to new distributions is still a fundamental challenge, especially when the distribution shifts are caused by unobserved confounded variables. An example is the data source platform (unobserved) may spurious correlated to the sentiment label during and this correlation maybe change at test time.

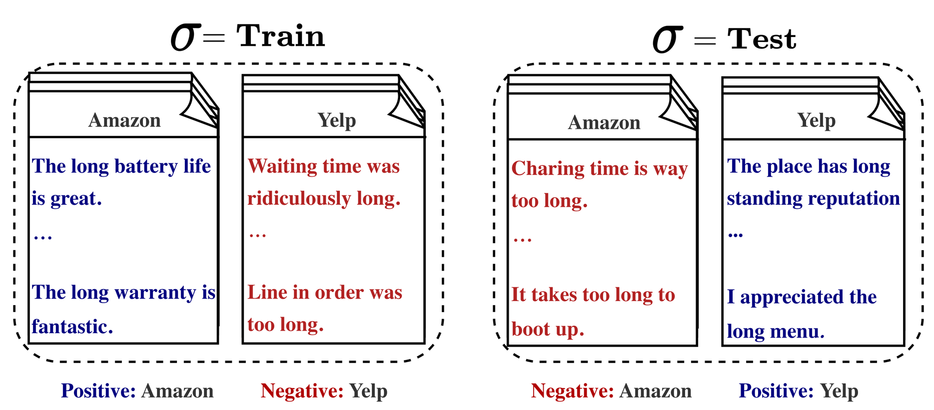

Fig. 1: Sentiment is associated with data source: Amazon with positive sentiment and Yelp with negative sentiment, which reverts in test regime.

Problem

- Can be formulated as a domain generalization problem. But often lack of clear definition on **what to generalize** from training data alone.
- Supervised fine-tuning surely fail, what else can we do to learn *meaningful causal structures*. How do we construct a **causally robust predictor** to automatically generalize to test time distribution?
- We formulate fine-tuning from pre-trained models as **causal identification problem**.

Literature

- Distribution shift is an ill-posed problem without assumptions [2]. Central assumptions on which part of data generative process is invariant, which can be answered with tools from causal transportability theory [5, 3].
- Most common assumptions on either covariate shift or label shift, later extend to causal motivated robust representation learning, i.e. learning an invariant $\Phi(X)$. Based on either multiple environments (e.g. IRM [1]) or counterfactual augmentation [4].
- Recent focus on causal representation learning methods such as learning stable causal latent variable [6], invariant predictor [7] and compositional models [9].
- Often requires multiple intervention data regimes or environment labels, which can be impractical. Our work build on compositional models approach and try to identify useful components under standard supervised learning setup, such that these components can adjust for test time adaptation.

Problem Statement

Given: Text X and Label Y with unobserved confounded variable U, where σ denotes the data generation regime. During training and test, the σ change indicating $p(U)$ change arbitrarily. The goal is to learn a classifier that are able to adapt to the changes.

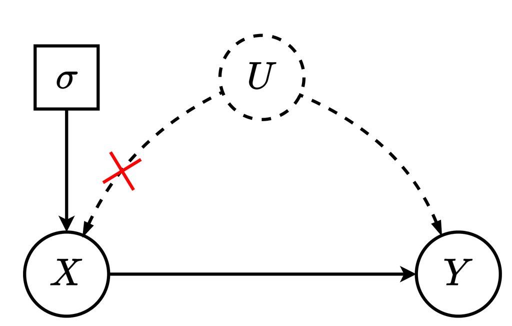

Fig. 2: Causal diagram for the problem.

Structural Assumptions

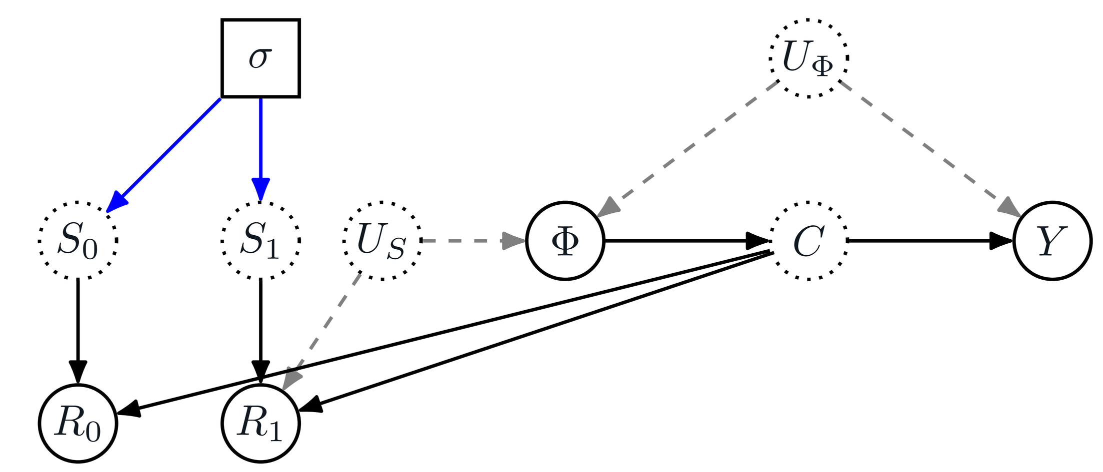

Fig. 3: Refinement of the original causal diagram, where X is broken apart and abstracted into vectors R_0, R_1 and Φ.

Theorem 0.1 (**Identification for Causal Features** C). *Assume the structural assumptions encoded in the causal graph in* ***Fig. 3****. Let the mapping between* $\{S_0, S_1, C\}$ *and* $\{R_0, R_1, \Phi\}$ *obey the invertibility conditions of [8]. According to* ***Theorem 4.4*** *in [8], we can identify* C *by learning the distribution* $p(c \mid r)$ *from* R_0 *and* R_1.

Intuition: We can identify the invariant latent variable when having access to more than one view of same stable variable with different generative process induced by none-stable variable.

Theorem 0.2 (**Identification for Causal Transfer Learning**). *Given the assumptions in the causal graph in* ***Fig. 3*** *and Theorem 0.1, the distribution of* Y *under* do(x) *can be computed as*[1]

$$p(y \mid \text{do}(x)) = \sum_{\Phi', x'} p(y \mid \Phi', c) p(\Phi' \mid x') p(x'), \tag{1}$$

where c *is given by* $c = p(c|r_1)$ *and* $r_1 = p(r_1|x)$. □

Intuition: We can perform causal fine-tuning by marginalization over the possible (but never observed in observational data) spurious distribution over the entire dataset.

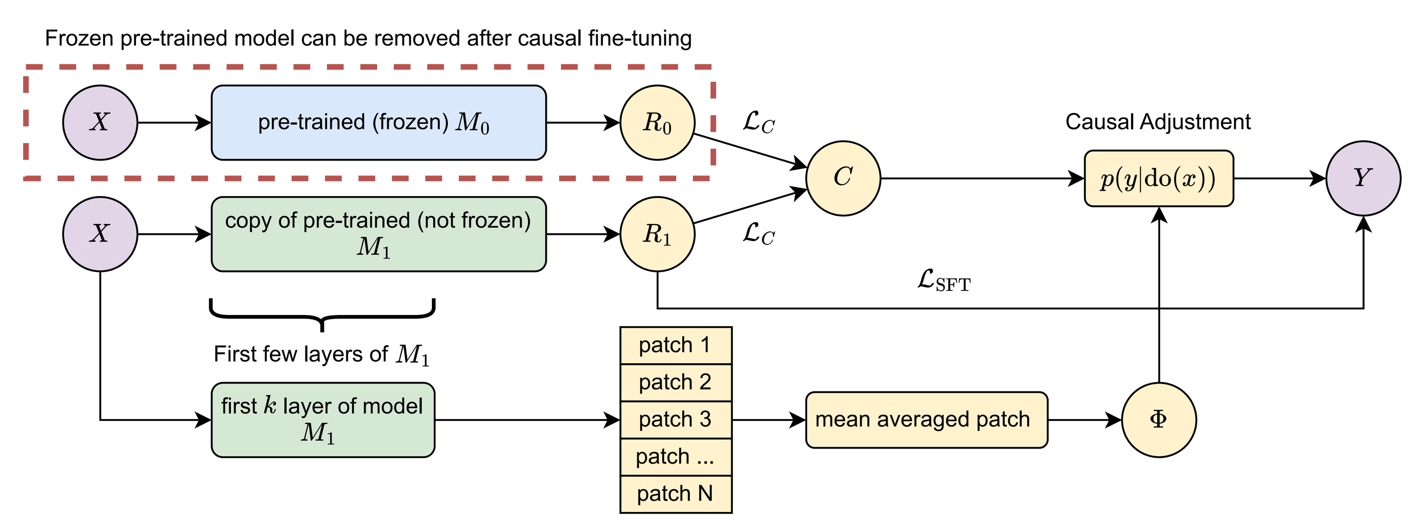

Fig. 4: Illustration for causal fine-tuning method.

Submodule 1: Supervised Fine-Tuning The first submodule learns $p(r_1 \mid x)$ from training samples of $p(x, y)$ through supervised fine-tuning (SFT) where $p(r_1 \mid x)$ is initialized with the pre-trained model $p(r_0 \mid x)$.

Submodule 2: Learning Causal Feature To learn the invariant causal feature C, we aim to identify the distribution $p(c \mid r)$. This process involves aligning representations from different environments while maximizing entropy to prevent collapsed representations [8].

Submodule 3: Retrieving Local Feature Given input X as a series of tokens $X = [t_1, t_2, ..., t_m]$, we can retrieve the vector representation for each token t at the embedding layers from the SFT model. To construct local feature Φ, we divide the token sequence into non-overlapping patches, allowing us to rewrite X as patches $X = [p_1, p_2, ..., p_{10}]$ where $p_1 = [t_1, t_2, ..., t_{\frac{m}{10}}]$ and so on. After splitting, we perform mean averaging on these patches to extract the local feature Φ, which is then used with C together to estimate $p(y \mid \text{do}(x))$.

Experiments

Data: 1. Semi-synthetic data: spurious correlation between stop words and label. and 2. Semi-synthetic data: spurious correlation between data source and label.

CFT Models: 1. CFT: proposed model. 2. CFT-N, CFT-C, CFT-Φ: ablation model.

Baselines: SFT0 and SFT.

Results

	Train F1 90%	ID F1 90%	OOD F1 70%	OOD F1 50%	OOD F1 30%	OOD F1 10%
SFT0	86.24	86.42	71.58	56.82	42.04	26.94
SFT	95.96	92.89	81.89	71.20	60.23	49.24
CFT	**98.69**	**93.03**	**84.16**	**75.83**	**67.06**	**58.40**
CFT-N	97.80	92.35	81.91	71.89	61.46	51.07
CFT-C	98.62	92.99	84.07	75.51	66.62	57.75
CFT-Φ	92.42	89.30	71.83	54.41	36.91	19.08

Fig. 5: Main experimental results, averaged over five different seeds.

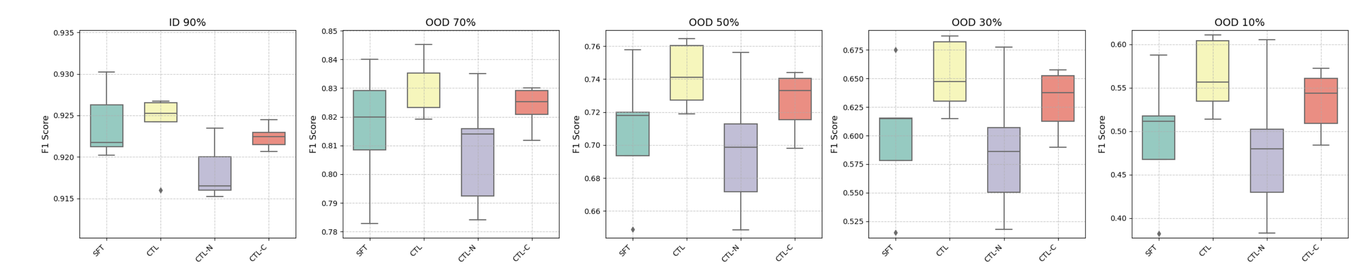

Fig. 6: Box-plot over 5 runs for 4 methods (SFT, CFT, CFT-N and CFT-C). Some methods are not included as they are significantly worse.

Conclusion

1. The results show the superiority of our model against strong baselines. 2. We also observed that the structural assumptions are critical for latent confounded shift and robust test time adaptation. 3. We show that pre-trained models can be utilized to train causal classifiers.

References

[1] Martin Arjovsky et al. "Invariant risk minimization". In: *arXiv preprint arXiv:1907.02893* (2019).

[2] Jiayuan Huang et al. "Correcting sample selection bias by unlabeled data". In: *Advances in neural information processing systems* 19 (2006).

[3] Kasra Jalaldoust and Elias Bareinboim. "Transportable representations for domain generalization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 11. 2024, pp. 12790–12800.

[4] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. "Learning The Difference That Makes A Difference With Counterfactually-Augmented Data". In: *International Conference on Learning Representations*.

[5] Judea Pearl and Elias Bareinboim. "Transportability of causal and statistical relations: A formal approach". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 25. 1. 2011, pp. 247–254.

[6] Xinwei Sun et al. "Recovering latent causal factor for generalization to distributional shifts". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 16846–16859.

[7] Victor Veitch et al. "Counterfactual invariance to spurious correlations in text classification". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 16196–16208.

[8] Julius Von Kügelgen et al. "Self-supervised learning with data augmentations provably isolates content from style". In: *Advances in neural information processing systems* 34 (2021), pp. 16451–16467.

[9] Jialin Yu et al. "Structured Learning of Compositional Sequential Interventions". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 115409–115439.

[1]