Towards the Causal Complete Cause of Multi-Modal Representation Learning

Jingyao Wang[1], Siyu Zhao[1], Wenwen Qiang*, Jiangmeng Li, Changwen Zheng, Fuchun Sun, Hui Xiong

Introduction

Multi-Modal Learning (MML) aims to learn effective representations across modalities for accurate predictions. Existing methods typically focus on modality consistency and specificity to learn effective representations. However, from a causal perspective, they may lead to representations that contain insufficient and unnecessary information. To address this, we propose that effective MML representations should be causally sufficient and necessary. Considering practical issues like spurious correlations and modality conflicts, we relax the exogeneity and monotonicity assumptions prevalent in prior works and explore the concepts specific to MML, i.e., Causal Complete Cause (C^3). We begin by defining C^3, which quantifies the probability of representations being causally sufficient and necessary. We then discuss the identifiability of C^3 and introduce an instrumental variable to support identifying C^3 with non-exogeneity and non-monotonicity. Building on this, we conduct the C^3 measurement, i.e., C^3 risk. We propose a twin network to estimate it through (i) the real-world branch: utilizing the instrumental variable for sufficiency, and (ii) the hypothetical-world branch: applying gradient-based counterfactual modeling for necessity. Theoretical analyses confirm its reliability. Based on these results, we propose C^3 Regularization, a plug-and-play method that enforces the causal completeness of the learned representations by minimizing C^3 risk. Extensive experiments demonstrate its effectiveness.

Contributions

1. We propose the definition, identifiability, and measurement of causal sufficiency and necessity, i.e., causal complete cause, for MML without exogeneity and monotonicity assumptions.
2. We theoretically demonstrate the effectiveness and reliability of the proposed measurement, i.e., C^3 risk, and propose C^3R, which can be applied to any MML model to learn causal complete representations with low C^3 risk.
3. We conduct extensive experiments on various datasets and multi-modal baselines that prove the effectiveness and robustness of C^3R.

Motivation and Analysis

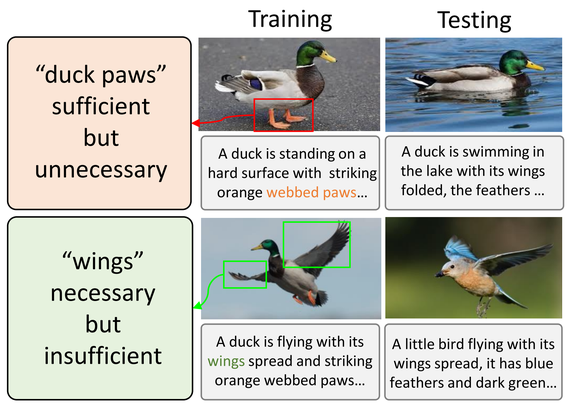

Figure 1: Example of causal sufficiency and necessity.

Human perception integrates diverse modalities such as vision, hearing, and touch. Multimodal learning (MML) mimics this by learning representations from multiple modalities for accurate prediction. Existing methods focus on *modality consistency* (aligning features across modalities) or *modality specificity* (preserving unique modality traits). However, from a causal view, such representations may be insufficient or unnecessary. We define *causal sufficiency* as the ability to predict labels from representations, and *causal necessity* as the prediction changing when representations are removed. Optimizing only one leads to poor generalization or spurious cues. Our experiments confirm this. We argue that effective MML representations must satisfy both—i.e., be *causally complete*. Yet, enforcing this is challenging: common assumptions like exogeneity and monotonicity often fail due to semantic entanglement, modality conflict, and nonlinear interactions. This limits the reliability of traditional causal constraints in real-world multimodal scenarios (The constructed SCMs illustrate this).

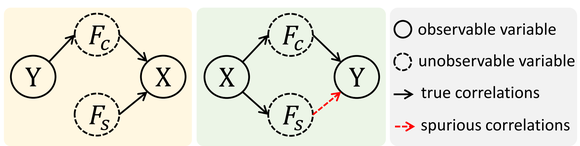

Figure 2: Structural Causal Model (SCM) for MML. **Left:** causal generating mechanism, **Right:** the learning process.

Methodology

To address these issues, we relax the traditional assumptions of exogeneity and monotonicity, and explore causal sufficiency and necessity in MML to ensure representation quality. We first formalize the concept of Causal Complete Cause (C^3), which quantifies the probability that label predictions (Y) change under two types of interventions on the representation (Z)—one assessing sufficiency, the other assessing necessity. We then analyze the identifiability of C^3 and introduce instrumental variables V to enable estimation from observational data, even under relaxed assumptions. Based on this, we propose a twin network to estimate the C^3 risk, where low risk indicates high-confidence causal completeness. This estimation faces two key challenges: eliminating spurious correlations in sufficiency assessment and generating counterfactuals for necessity evaluation. The twin network addresses these via (i) a real-world branch that removes spurious effects using instrumental variables, and (ii) a counterfactual branch that constructs counterfactuals through provable gradient-based perturbation. We introduce C^3 Regularization (C^3R), a plug-and-play training strategy that learns causally complete multimodal representations by minimizing C^3 risk.

$$\min_{f_\theta} \widehat{R}^{C^3} + \lambda_v \mathcal{L}_v + \lambda_{fe} \mathcal{L}_{fe} \tag{1}$$

Theoretical analysis establishes its reliability and provides performance guarantees for the C^3 risk.

Experimental Results

We conduct extensive experiments on various downstream tasks, and the results show that the introduction of C^3R achieves stable performance improvements on multiple baselines in both the average and worst-case accuracy. In addition, the results of visualization experiments and corner cases analyses on multiple benchmark datasets further demonstrate the effectiveness of C^3 in learning causal complete causes.

Table 1: Performance comparison when 50% samples are corrupted with Gaussian noise, i.e., zero mean with the variance of N. "(N, Avg.)" and "(N, Worst.)" denotes the average and worst-case accuracy. The best results are highlighted in **bold**.

Method	NYU Depth V2				SUN RGB-D				FOOD 101				MVSA			
	(0,Avg.)	(0,Worst.)	(10,Avg.)	(10,Worst.)	(0,Avg.)	(0,Worst.)	(10,Avg.)	(10,Worst.)	(0,Avg.)	(0,Worst.)	(10,Avg.)	(10,Worst.)	(0,Avg.)	(0,Worst.)	(10,Avg.)	(10,Worst.)
CLIP (Sun et al., 2023)	69.32	68.29	51.67	48.54	56.24	54.73	35.65	32.76	85.24	84.20	52.12	49.31	62.48	61.22	31.64	28.27
ALIGN (Jia et al., 2021)	66.43	64.33	45.24	42.42	57.32	56.26	38.43	35.13	86.14	85.00	53.21	50.85	63.25	62.69	30.55	26.44
MaPLe (Khattak et al., 2023)	71.26	69.27	52.98	48.73	62.44	61.76	34.51	30.29	90.40	86.28	53.16	40.21	77.43	75.36	43.72	38.82
CoOp (Jia et al., 2022a)	67.48	66.94	49.43	45.62	58.36	56.31	39.67	35.43	88.33	85.10	55.24	51.01	74.26	73.61	42.58	37.29
VPT (Jia et al., 2022a)	62.16	61.21	41.05	37.81	54.72	53.92	33.48	29.81	83.89	82.00	51.44	49.01	65.87	64.98	32.79	29.21
Late fusion (Wang et al., 2016)	69.14	68.35	51.99	44.95	62.09	60.55	47.33	44.60	90.69	90.58	58.00	55.77	76.88	74.76	55.16	47.78
ConcatMML (Zhang et al., 2021)	70.30	69.42	53.20	47.71	61.90	61.19	45.64	42.95	89.43	88.79	56.02	54.33	75.42	75.33	53.42	50.47
AlignMML (Wang et al., 2016)	70.31	68.50	51.74	44.19	61.12	60.12	44.19	38.12	88.26	88.11	55.47	52.76	74.91	72.97	52.71	47.03
ConcatBow (Zhang et al., 2023c)	49.64	48.66	31.43	29.87	41.25	40.54	26.76	24.27	70.77	70.68	35.68	34.92	64.09	62.04	45.40	40.95
ConcatBERT (Zhang et al., 2023c)	70.56	69.83	44.52	43.29	59.76	58.92	45.85	41.76	88.20	87.81	49.86	47.79	65.59	64.74	46.12	41.81
MMTM (Joze et al., 2020)	71.04	70.18	52.28	46.18	61.72	60.94	46.03	44.28	89.75	89.43	57.91	54.98	74.24	73.55	54.63	49.72
TMC (Han et al., 2020)	71.06	69.57	53.36	49.23	60.68	60.31	45.66	41.60	89.86	89.80	61.37	61.10	74.88	71.10	60.36	53.37
LCKD (Wang et al., 2023b)	68.01	66.15	42.31	40.56	56.43	56.32	43.21	42.43	85.32	84.26	47.43	44.22	62.44	62.27	43.52	38.63
UniCODE (Xia et al., 2024)	70.12	68.74	44.78	42.79	59.21	58.55	46.32	42.21	88.39	87.21	51.28	47.95	66.97	65.94	48.34	42.95
SimMMDG (Dong et al., 2024)	71.34	70.29	45.67	44.83	60.54	60.31	47.86	45.79	89.57	88.43	52.55	50.31	67.08	66.35	49.52	44.01
MMBT (Kiela et al., 2019)	67.00	65.84	49.59	47.24	56.91	56.18	43.28	39.46	91.52	91.38	56.75	56.21	78.50	78.04	55.35	52.22
QMF (Zhang et al., 2023c)	70.09	68.81	55.60	51.07	62.09	61.30	48.58	47.50	92.92	92.72	62.21	61.76	78.07	76.30	61.28	57.61
CLIP+C^3R	76.54	**75.12**	56.73	52.90	62.31	58.71	41.59	37.52	92.93	91.80	59.77	57.54	69.61	68.64	39.58	35.89
MaPLe+C^3R	77.07	74.45	58.94	55.95	66.21	65.51	40.12	37.34	94.38	93.51	60.63	46.07	81.19	81.51	49.32	45.98
Late fusion+C^3R	73.26	71.62	57.21	50.98	64.84	63.25	**53.35**	50.43	94.09	92.24	65.27	59.02	**83.77**	79.79	62.14	52.50
LCKD+C^3R	77.14	75.12	50.11	47.98	60.97	60.14	47.23	46.21	90.89	90.14	54.48	51.16	66.78	65.67	49.28	42.84
SimMMDG+C^3R	75.32	74.61	49.99	47.22	65.50	64.58	52.69	**51.70**	92.24	91.14	57.32	53.56	73.62	71.01	51.65	51.07
MMBT+C^3R	73.74	71.82	54.35	52.57	61.47	59.99	48.42	46.07	94.25	93.90	60.41	60.11	82.76	**81.64**	62.12	58.93
QMF+C^3R	**77.58**	74.95	**59.72**	**59.18**	**67.35**	**65.84**	52.26	51.28	**94.87**	**93.79**	**66.45**	**63.69**	83.13	81.98	**66.66**	**64.51**

Resources

Website | CODE | Paper | ICML Page

References

[1] Judea Pearl. *Causality*. Cambridge university press, 2009.