

Learning Invariant Causal Mechanism from Vision-Language Models

宋泽恩*, 赵思雨*, 张星宇*, 李江梦, 郑昌文, 强文文

International Conference on Machine Learning

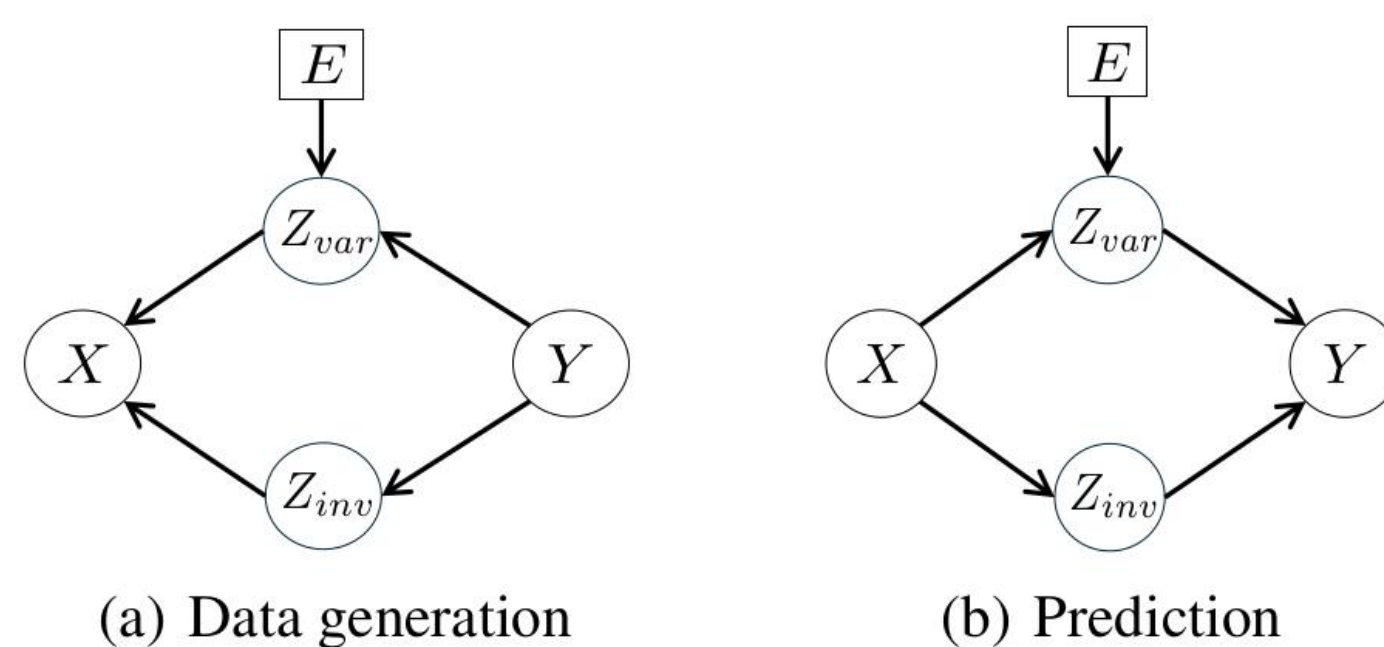
宋泽恩 zeensong@qq.com

概述

视觉-语言模型的代表应在因果意义上具有环境不变性以实现稳健泛化。为解决 CLIP 在分布外环境下性能退化的问题，我们提出 CLIP 不变因果机制 (CLIP-ICM)，并在证明 CLIP 表征可线性分解为不变与变异因子基础上，引入干预数据实现不变子空间的可识别性。基于此，设计三阶段框架，通过线性投影估计不变表征，并在不重训练主干网络的情况下进行不变预测。理论分析与实验证明，CLIP-ICM 有效提升了 CLIP 的因果鲁棒性与 OOD 泛化性能。

动机与分析

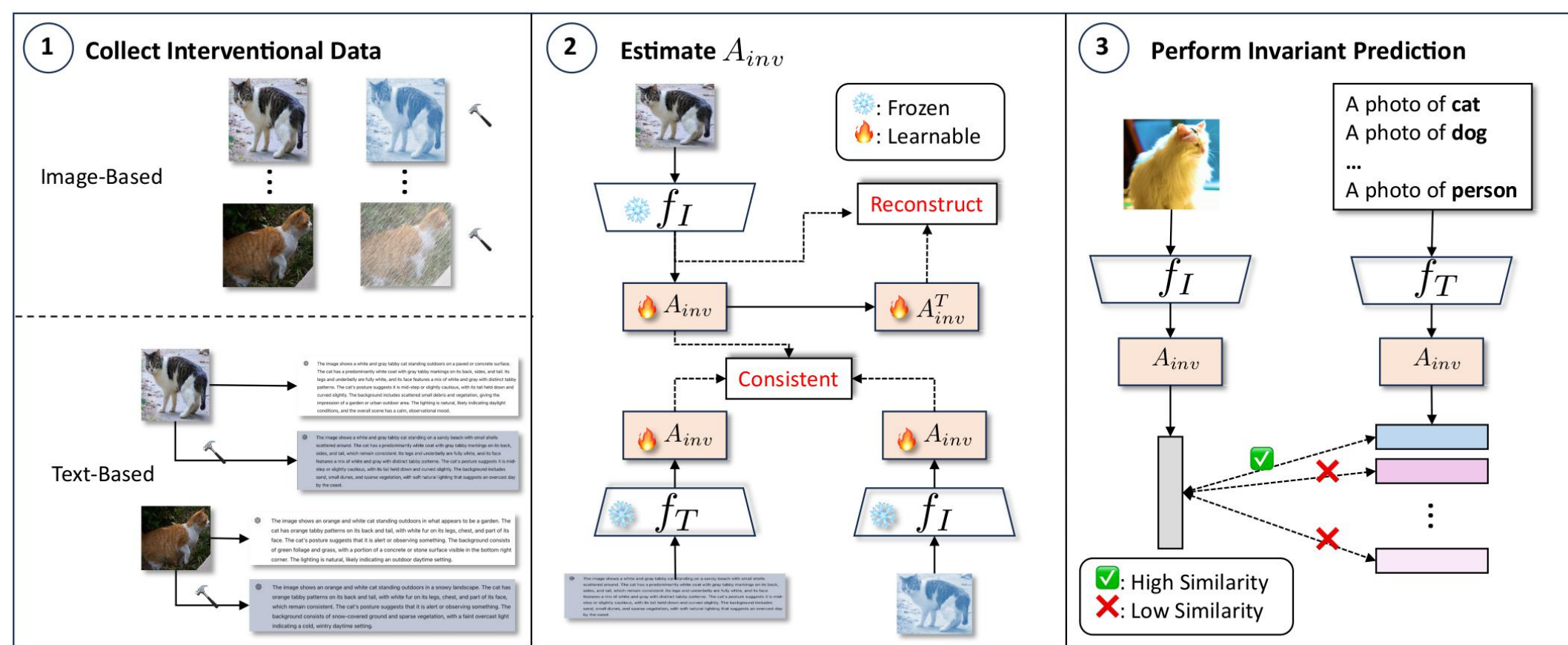
人类认知依赖捕捉环境不变的因果规律，而视觉-语言模型常因混合环境不变与变异因素，导致分布外 (OOD) 性能退化。现有方法未充分保证预测机制的因果一致性，易引入虚假关联或丢失关键信息。我们认为，稳健的表征应具备因果不变性，即在不同环境下保持一致的因果机制。为此，我们基于结构因果模型 (SCM) 分析 CLIP 表征的可分解性，并探讨如何在重训练主干网络的情况下提取不变因子，提升 OOD 泛化。



图一 该图展示视觉-语言建模中的结构因果模型，其中 (a) 表示数据生成过程：环境变量 E 会影响环境相关的潜变量 Z_{var} ，而 Z_{var} 与环境不变的潜变量 Z_{inv} 共同生成观测图像 X ，且两者均受到类别标签 Y 的影响；(b) 表示预测过程：观测图像 X 被编码为 Z_{inv} 和 Z_{var} ，这两部分潜变量共同决定最终的类别预测 Y 。

方法

我们提出 CLIP 不变因果机制 框架，旨在从 CLIP 表征中提取环境不变因素。具体做法是，首先收集不同环境下的干预数据，利用这些数据揭示不变与变异因素的差异。接着，我们在理论上证明 CLIP 表征可以线性分解为不变和可变成分，据此学习线性投影矩阵，将图像和文本映射到统一的不变子空间。最后，在该空间内进行预测，确保依赖一致的因果机制。



图二 CLIP-ICM 方法流程。① 收集干预数据区分环境因素；② 学习投影矩阵提取不变特征；③ 在不变空间内进行稳健预测。

实验

在多个典型 OOD 基准数据集上的系统评估结果表明，该方法在 Terra Incognita、PACS、Office Home、DomainNet 等挑战性测试集中均取得显著性能提升，超越现有主流微调方法。在保证主干冻结的前提下，该方法不仅实现了更高的准确率，还展现出更强的环境鲁棒性与类别迁移能力。

METHOD	BACKBONE	PACS	VLCS	OFFICEHOME	TERRAINC	DOMAINNET	AVG.
ZERO-SHOT	CLIP	96.1	82.4	71.5	34.2	56.8	68.2
LINEAR-PROBE	CLIP	96.4±0.1	78.7±0.2	81.9±0.4	60.2±0.2	55.0±0.4	74.4±0.4
MIRO (CHA ET AL., 2022)	CLIP	95.6±0.2	82.2±0.1	82.5±0.3	54.3±0.2	54.0±0.5	73.7±0.5
COOP (ZHOU ET AL., 2022B)	CLIP	97.0±0.2	83.0±0.1	81.1±0.4	54.6±0.2	59.5±0.2	75.0±0.2
CoCoOP (ZHOU ET AL., 2022A)	CLIP	96.7±0.4	83.6±0.1	80.7±0.1	56.2±0.3	59.7±0.4	75.4±0.4
CLIP-ADAPTER (GAO ET AL., 2023)	CLIP	96.4±0.3	84.3±0.5	82.2±0.2	57.5±0.4	59.9±0.1	76.1±0.1
DPL (ZHANG ET AL., 2022)	CLIP	97.3±0.5	84.3±0.1	84.2±0.2	52.6±0.4	56.7±0.4	75.0±0.4
CLIPOOD	CLIP	97.3±0.1	85.0±0.4	87.0±0.2	60.4±0.7	63.5±0.1	78.6±0.1
CLIP-ICM*	CLIP	97.3±0.5	84.1±0.4	82.6±0.3	49.9±0.3	60.5±0.3	74.9±0.4
CLIP-ICM* LINEAR-PROBE	CLIP	97.5±0.5	86.5±0.1	84.6±0.4	64.3±0.3	64.0±0.2	79.0±0.3
CLIP-ICM†	CLIP	96.8±0.4	83.4±0.3	82.1±0.4	45.2±0.3	57.4±0.2	73.0±0.3
CLIP-ICM† LINEAR-PROBE	CLIP	97.2±0.5	85.2±0.5	82.4±0.3	61.2±0.1	59.6±0.4	77.1±0.4
CLIP-ICM	CLIP	97.7±0.2	86.2±0.3	84.6±0.2	52.5±0.4	61.1±0.3	76.4±0.3
CLIP-ICM LINEAR-PROBE	CLIP	97.8±0.3	86.6±0.1	87.1±0.4	66.5±0.1	65.0±0.1	80.6±0.2

表一 DomainBed 基准上在域移场景下的准确率。标注 * 的方法仅使用基于图像的干预数据训练，标注 † 的方法仅使用基于文本的干预数据训练。

SPLIT	METHOD	OFFICEHOME				DOMAINNET					
		A	C	P	R	C	I	P	Q	R	S
BASE	CLIP	86.8	75.5	89.5	92.6	72.8	51.7	66.0	13.5	83.4	66.9
	CoOp	87.0±0.4	78.3±1.2	92.4±0.2	91.4±0.6	75.7±0.2	58.8±0.5	68.5±1.3	13.1±1.0	84.0±0.5	70.0±0.1
	CLIPOOD	90.1±0.2	79.7±0.2	93.1±0.1	94.8±0.1	79.0±0.2	62.2±0.1	73.0±0.2	20.2±0.2	86.2±0.1	73.8±0.1
	CLIP-ICM*	88.6±0.4	78.0±0.2	90.2±0.3	93.1±0.2	74.4±0.3	53.2±0.3	67.2±0.3	14.6±0.2	85.2±0.3	67.8±0.2
	CLIP-ICM†	87.1±0.1	77.2±0.2	89.3±0.5	92.0±0.2	73.1±0.2	52.0±0.1	66.1±0.1	14.1±0.1	83.8±0.4	66.8±0.4
	CLIP-ICM	89.2±0.4	78.6±0.4	90.6±0.1	93.7±0.2	75.0±0.3	53.9±0.3	67.6±0.4	14.9±0.3	85.9±0.3	68.4±0.4
NEW	CLIP	76.6	59.4	88.1	86.2	70.2	44.1	66.4	14.1	83.5	61.0
	CoOp	76.5±1.1	56.6±2.4	88.0±1.9	86.8±0.7	71.5±0.2	47.2±0.3	67.3±0.7	14.8±0.7	83.7±0.7	63.0±0.3
	CLIPOOD	77.8±0.2	60.0±0.2	88.3±0.1	86.7±0.1	71.2±0.1	48.1±0.1	68.2±0.2	18.0±0.4	83.4±0.1	62.9±0.1
	CLIP-ICM*	81.7±0.4	66.5±0.5	90.2±0.4	90.6±0.4	76.7±0.2	50.9±0.2	69.1±0.5	17.2±0.5	83.6±0.3	67.7±0.5
	CLIP-ICM†	81.2±0.2	65.2±0.2	89.4±0.4	89.9±0.2	76.0±0.2	49.8±0.1	67.9±0.2	15.7±0.1	82.5±0.1	67.0±0.4
	CLIP-ICM	82.6±0.1	67.5±0.2	90.9±0.3	91.5±0.3	77.8±0.1	51.6±0.5	70.2±0.2	18.0±0.4	84.5±0.2	68.6±0.5
TOTAL	CLIP	82.6	67.3	88.8	89.5	71.4	47.1	66.2	13.8	83.4	63.4
	CoOp	82.7±0.5	67.2±0.7	90.2±1.0	89.2±0.6	73.4±0.3	51.8±0.3	67.9±1.0	13.7±0.8	83.9±0.5	66.0±0.2
	CLIPOOD	85.1±0.1	69.6±0.2	90.8±0.1	91.0±0.1	74.8±0.1	53.6±0.1	70.6±0.1	19.1±0.3	84.8±0.1	67.4±0.1
	CLIP-ICM*	85.2±0.4	71.2±0.3	89.4±0.3	91.9±0.3	75.6±0.2	52.1±0.2	68.2±0.4	15.9±0.3	84.4±0.3	67.8±0.3
	CLIP-ICM†	84.2±0.1	71.2±0.2	89.4±0.5	91.0±0.2	74.6±0.2	50.9±0.1	67.0±0.1	14.9±0.1	83.2±0.2	66.9±0.4
	CLIP-ICM	85.9±0.2	73.1±0.3	90.8±0.2	92.6±0.2	76.4±0.2	52.8±0.4	68.9±0.3	16.5±0.3	85.2±0.2	68.5±0.5

表二 OfficeHome 和 DomainNet 数据集上同时存在域移和开放类场景下的准确率。标注 * 的方法仅使用基于图像的干预数据训练，标注 † 的方法仅使用基于文本的干预数据训练。