计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(7):1−13 [doi: 10.15888/j.cnki.csa.009942] [CSTR: 32024.14.csa.009942]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

因果解耦表征学习综述①

黄贝贝, 刘进锋

(宁夏大学 信息工程学院, 银川 750021)
通信作者: 刘进锋, E-mail: jfliu@nxu.edu.cn

摘　要: 人工智能若想从根本上理解我们周围的世界, 关键在于它能否学会从所观察到的低级感官数据中识别并解开隐藏的潜在可解释因素. 解耦表征学习正是为了从数据中提取出这些独立且可解释的潜在变量, 而因果解耦表征学习则更进一步强调了这些潜在变量之间的因果关系, 从而更真实地模拟现实世界的复杂性. 鉴于因果学习的重要性日益增长, 本文对结合因果学习的解耦表征学习的相关方法进行了详细、全面地介绍, 旨在为解耦表征学习的未来发展提供支持. 根据常用的因果学习的相关方法对因果解耦表征学习进行分类, 主要探讨了结合结构因果模型和基于流模型的解耦表征学习方法以及常用的数据集与评价指标. 此外, 还分析了因果解耦表征学习在图像生成、3D 姿态估计和无监督领域适应等应用的实际案例, 并对未来的研究方向进行前瞻性展望, 为科研人员和实践者揭示未来可能的探索方向, 促进该领域的持续发展和创新.

关键词: 解耦表征学习; 因果关系; 结构因果模型; 流模型; 图像生成

引用格式: 黄贝贝,刘进锋.因果解耦表征学习综述.计算机系统应用,2025,34(7):1–13. http://www.c-s-a.org.cn/1003-3254/9942.html

Survey on Causally Disentangled Representation Learning

HUANG Bei-Bei, LIU Jin-Feng

(School of Information Engineering, Ningxia University, Yinchuan 750021, China)

Abstract: The key for artificial intelligence to fundamentally comprehend the world around us is to identify and disentangle hidden, potentially interpretable factors from observed low-level sensory data. Disentangled representation learning aims to extract these independent and interpretable latent variables from data, while causally disentangled representation learning further emphasizes the causal relationships among these latent variables, thereby more truly simulating the complexity of the real world. In light of the increasing importance of causal learning, this study provides a detailed and comprehensive introduction to relevant methods combining causal learning with disentangled representation learning, intending to support future development in disentangled representation learning. The study classifies causally disentangled representation learning based on commonly used causal learning methods, mainly discussing methods that integrate structural causal models with flow-based disentangled representation learning, as well as commonly used datasets and evaluation metrics. Furthermore, it analyzes practical applications of causally disentangled representation learning in image generation, 3D pose estimation, and unsupervised domain adaptation, and provides a forward-looking perspective on future research directions. This study reveals potential exploration paths for researchers and practitioners, promoting continuous development and innovation in this field.

Key words: disentangled representation learning; causal relationship; structural causal model; flow-based model; image generation

① 基金项目: 宁夏自然科学基金 (2023AAC03126)
收稿时间: 2024-12-13; 修改时间: 2025-01-07, 2025-02-11; 采用时间: 2025-03-06; csa 在线出版时间: 2025-05-23
CNKI 网络首发时间: 2025-05-26

近 10 年来, 随着机器学习——尤其是深度学习的发展, 人工智能的热潮席卷全球, 相关的研究及应用也在广泛展开, 但这表明繁荣的背后隐藏着巨大的危机. 深度学习的进步主要依赖于蛮力工程方法, 即主要依靠不断增大的模型以及数据来提升性能. 但其背后的技术实质是一种“黑盒”方法, 目前深度学习主要基于统计相关来学习, 模型学到的特征表示在很大程度上是隐藏的, 并且可解释性差. 最终学习到的可能不是事物的本质特征, 对没见过的情景或对抗性攻击缺乏稳定性; 另外, 学习到的很可能是相关关系而非因果关系, 相关关系可以用来作预测, 但只有因果关系才适合进行决策. 这导致深度学习在关键应用场景, 比如无人驾驶、医疗健康、工业制造、金融和司法等领域无法深入, 只能作为人类的一种辅助工具. 因此, 如何提高模型的可解释性和稳定性已成为亟待解决的核心问题.

在这种背景下, 解耦表征学习应运而生. 解耦表征学习通过挖掘数据中潜在的相互作用因子, 并赋予其相互分离的数据表征, 属于可解释性的深度表征学习范畴, 能够很大程度上提高深度学习的可解释性, 增强其内在逻辑性. 2013 年 Bengio 等人[1]提出关于解耦表征的直观定义: 解耦表征应分离数据中不同、独立和信息丰富的变异生成因素. 单一潜变量对单一基本生成因素的变化敏感, 而对其他因素的变化相对不变.

解耦表征学习不仅是对数据特征的简单提取, 而是对数据生成机制的深层次挖掘, 识别和分离数据中独立且具有因果关系的潜在因子. 这种能力使得模型能够模拟人类的认知过程, 即从复杂的信息中提取出有意义的模式和规律[2]. 根据不同的假设框架, 解耦表征学习可以分为基于独立假设和基于因果假设的两种类型. 因果解耦表征学习方法可以捕捉数据生成过程的潜在因果机制, 并通过分解因果因素实现更可解释和更稳健的表示. 与传统的统计学习方法相比, 因果解耦表征学习能够从数据中识别出因果关系, 从而超越仅依赖于观察数据分布和训练任务的局限, 具备更强的推理能力. 尽管统计学习已取得了诸多成果, 但它对现实世界的描述仍然较为片面, 需要在特定的实验条件下才能发挥作用. 相反, 因果学习试图结合数据驱动的学习与未包含在统计模型中的假设, 模拟干预和分布变化对系统的影响, 从而提供更加深刻的理解[3]. 这种结合因果推理与深度学习的方法, 为解决深度学习模型的可解释性和稳定性问题提供了新的思路和方法.

随着因果推理与深度学习的结合逐渐成为研究热点, 越来越多的学者开始关注如何通过因果学习提升深度学习模型的可解释性和稳健性. 这一趋势不仅反映了学术界对深度学习“黑盒”特性的普遍担忧, 也揭示了对模型在复杂现实世界应用的更高要求. 近年来, 相关研究不断增加, 特别是在无人驾驶、医疗诊断、金融风险评估等关键领域, 因果关系的精准识别以及决策过程的透明化已成为实现高可靠性系统的核心. 这些领域迫切需要能够提供清晰、可解释的决策过程的模型, 从而增强用户对模型的信任和接受度.

然而, 尽管解耦表征学习近年来逐渐得到研究者的关注, 但目前尚未发现有关于因果学习与解耦表征学习的研究进展综述发表. 这表明, 解耦表征学习与因果推理的结合仍是一个较为新颖且有待深入探讨的研究领域. 因此, 本研究旨在填补这一空白, 全面综述该领域的研究进展, 以揭示未来可能的探索方向, 促进该领域的持续发展和创新. 具体内容安排如下: 第 1 节介绍解耦表征学习的基本原理及相关研究工作; 第 2 节介绍因果学习并讨论因果学习的相关方法; 第 3 节综述因果解耦表征学习的研究工作; 第 4 节探讨因果解耦表征学习的应用领域; 第 5 节展望因果解耦表征学习的未来研究方向; 第 6 节对全文进行总结与归纳.

1 解耦表征学习

在解耦表征学习领域, 面对的问题是从一组高维观测数据中提取出独立且有意义的潜在因子. 具体来说, 假设有一组观测数据 $x \in R^D$, 其中 D 是数据的维度. 数据 x 由多个潜在因子 $z \in R^K$ 生成, 其中 K 是潜在因子的维度. 目标是从数据 x 中学习一个表示 z, 使得 z 中的各个因子是解耦的, 并且能够解释数据的生成过程.

为了实现这一目标, 通常使用生成模型来实现. 在这个框架下, 数据 x 被视为由潜在变量 z 通过某个生成分布生成的[4], 且在条件概率下, 可以将数据的生成过程表示为:

$$p(x|z) = Decoder(z) \tag{1}$$

其中, $p(x|z)$ 是数据在给定潜在因子 z 下的生成分布, $Decoder$ 是从潜在空间到观测空间的映射.

为了学习潜在因子的分布, 通常假设 $z \sim N(0, 1)$ (标准正态分布). 因此, 生成模型的目标是学习给定观测数据 x 时, 潜在因子 z 的后验分布:

$$p(z|x)=\frac{p(x|z)p(z)}{p(x)} \tag{2}$$

其中, $p(z)$ 是潜在因子的先验分布, $p(x)=\int p(x|z)p(z)\mathrm{d}z$ 是数据的边际似然.

解耦表征的核心在于独立性约束, 即潜在因子 z 中的各个分量 $z_1,z_2,\cdots,z_i$ 是独立的, 它们的联合分布可以分解为各个因子分布的乘积:

$$p(z)=\prod_{i=1}^{K}p(z_i) \tag{3}$$

为了确保这一点, 解耦表征学习通常通过引入正则化项来限制潜在因子之间的相关性. 具体地, 可以通过最小化潜在变量的互信息来鼓励解耦. 例如, 可以在模型的损失函数中增加一个正则化项, 确保潜在变量之间的互信息 $I(z_i, z_j)$ 尽可能小. 通常, 互信息 $I(z_i, z_j)$ 表示两个潜在因子 z_i 和 z_j 之间的相关性, 最小化互信息有助于促进更好地解耦. 如果解耦学习成功, $I(z_i, z_j)$ 会接近于 0, 意味着不同因子和之间的相关性被有效去除, 从而实现独立性.

总的来说, 解耦表征学习的目标是从复杂的观测数据中提取出独立且有意义的潜在因子, 为各个领域的应用做铺垫. 并且, 解耦表征学习采用了多种不同的方法和技术. 最早的解耦表征学习方法是基于独立成分分析 (independent component algorithm, ICA)[5]. ICA 通过将测量信号表示为独立成分的线性组合, 试图从多变量信号中分离出统计独立的源, 然而现实世界中的数据往往具有统计相关性, 这与 ICA 所依赖的基本假设 (即信号源相互独立) 相悖. 由于 ICA 的局限性, 研究者们开始寻求其他更好的方法. 在这些方法中, 变分自编码器 (variational auto-encoder, VAE)[6]成为解耦表征学习领域中一个极具代表性的框架, 该模型通过引入不同的正则化项和先验分布来控制潜在空间的结构, 使得不同的潜在因子能够独立地表示数据的不同变化因子.

随着研究的不断深入, 众多解耦表征学习方法在 VAE 的基础上进行了扩展, 通过引入归纳偏好和采用多种正则化技术来实现更有效的解耦. 例如, Higgins 等人[7]提出的 β-VAE 框架最大化生成真实数据的概率以及最小化真实后验分布和估计后验分布的 KL 散度, 鼓励模型学习更有效的数据生成因素的可解释分解表示. β-TCVAE[8]在 β-VAE 的基础上, 惩罚总相关性, 用于学习解耦表征. Locatello 等人[9]证明了在没有归纳偏好的情况下, 无监督解耦学习理论上是不可能的. Locatello 等人[10]提出了通过引入监督信息的方式来改进解耦学习, 使用少量标签信息或者通过对比不同观测值的差异来帮助模型进行更精确的解耦. Sanchez 等人[11]于 2020 年提出一个基于互信息估计的模型, 不依赖于图像重建或图像生成, 通过最大化互信息来捕获数据属性, 同时运用对抗思想最小化共享和独有表示之间的互信息以强制表示解耦.

然而, 传统的解耦表征学习方法主要侧重于潜在因子之间的独立性, 往往忽略了数据中潜在的因果结构. 因果解耦表征学习则在解耦表征学习的基础上, 引入了因果关系的考量. 它不仅关注潜在因子之间的独立性, 更强调潜在因子与数据生成过程中的因果关系之间的联系. 理想情况下, 潜在因子 z 中的每个因子应该与数据生成过程中的一个独立因果因子一一对应, 并通过这些因子可以推导出数据的因果结构. 通过引入因果结构, 因果解耦表征学习使得模型不仅能够从数据中提取独立因子, 还能够推断数据生成过程中的因果关系. 这种因果推理的能力, 不仅提高了模型的可解释性, 还使得模型能够在面对新环境或任务时, 进行更有效的推理和决策, 在解决实际问题时提供更有价值的洞见和支持.

2 因果学习

2.1 因果学习的概述

因果学习作为一个日益重要的研究领域, 已经开始受到学术界的广泛关注. 通过建立因果图、因果推断算法等工具, 模型能够更精准地揭示变量之间的因果关系, 使得模型的决策过程不再是一个“黑箱”, 提供更透明、可解释的决策. 它不仅是一种技术手段, 更是一种哲学思考, 它要求我们深入探究数据背后的逻辑链条, 识别变量之间的因果联系, 为模型的决策提供更为精确和细致的解释.

因果学习与相关性分析有着本质的不同. 相关性分析通常揭示的是变量间的统计关系, 反映的是变量在同一时间段内是否一起变化, 但它并不考虑因果关系的方向性和产生这种关系的根本原因. 比如, 冰激凌销量和溺水事故之间确实存在相关性, 但这并不意味着冰激凌消费直接导致溺水事故的发生. 因果学习则试图明确一个变量是否为另一个变量变化的根本原因, 并关注潜在的混杂因素、外部干预效应以及实验设计

等关键要素. 通过因果学习, 明确因果关系背后的因果链条, 从复杂的数据中提炼出有意义的因果信息[12].

在解耦表征学习中, 因果学习起着至关重要的作用. 许多传统解耦方法可能会受到数据表面相关性的干扰, 无法准确识别哪些因子是关键的生成因素, 哪些是无关的噪声. 因果推断能够帮助模型区分这些因子, 识别出对数据生成过程有实质性影响的因子, 从而避免了单纯依赖相关性带来的误区. 通过引入因果学习, 解耦表征学习不仅能够更好地理解数据的内在结构, 还能在更复杂的环境中做出更为精确的预测和干预模拟. 这对于需要处理复杂因果关系的任务, 例如医疗干预、社会科学研究等领域, 具有重要的实际应用价值.

2.2 因果学习的相关方法

2.2.1 结构因果模型

结构因果模型 (structural causal model, SCM)[13] 在农业、社会科学和计量经济学等领域已经使用了很长时间, 是将因果关系和概率陈述联系起来的重要工具. 在因果解耦表征学习中, 结构因果模型通过三元组进行定义 $M=\langle Z,E,F\rangle$. 其中 Z 表示由 n 个内生因果变量 $z_1,z_2,\cdots,z_n$ 构成的集合; E 表示由 n 个外生噪声变量 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 构成的集合, 这些噪声变量通常作为中间潜在变量进行学习; F 由 n 个独立的因果模型构成的集合, 每个因果关系模型的形式为 $z_i=f_i(\varepsilon_i;z_p^i)$, 其中 f_i 是一个函数, 将噪声变量 ε_i 和父变量 z_p^i 映射到因果变量.

为了可以直观地表示因果关系, 通常采用有向无环图 (directed acyclic graph, DAG) 来展示变量间的因果结构.

结构因果模型的框架可以用来封装因果知识, 以及回答干预性和反事实的查询[14]. SCM 是基于因果关系的数据生成模型, 通过结构方程和有向图来表达因果效应, 可以有效进行因果建模和推断.

2.2.2 独立因果机制

统计学习通常基于从某一分布中采样的数据. 从因果建模的角度来看, 数据分布的结构并非偶然, 而是由背后的物理因果机制决定. 这些机制在生成数据的过程直接影响着不同观测变量之间的依赖关系. 因果机制可以被看作是生成模型中的独立模块, 这些模块不仅局限于特定的数据分布, 而且具备更强的跨领域适用性. 这意味着, 学习到的因果机制可以在不同任务和数据领域之间重复使用, 从而实现高效的迁移学习.

独立因果机制 (independent causal mechanisms, ICM)[15]是建立在因果机制模块化和独立性的基础之上的. 每个因果机制负责处理数据中的某一特定变换, 并且这些变换之间是独立的, 不相互干扰. 这种独立性使得因果机制能够在多个领域或任务之间迁移, 而无需重新学习每一个新的任务的细节. 通过模块化和独立性因果机制能够在处理数据时专注于数据的某一部分, 进行独立的因果操作. 这种结构使得因果机制在面对不同的数据变换时, 能够保持鲁棒性, 灵活适应新的数据分布, 同时保持因果解耦的效果.

在因果解耦表征学习中, 传统方法通常假设潜在因子是独立的, 这种假设在实际应用中往往是不现实的, 因为数据包含复杂的因果依赖关系, 这使得简单的独立性假设不再适用. 独立因果机制通过明确识别并建模这些因果关系, 不仅为因果推断提供了有效的工具, 也为跨任务学习和因果建模提供了强有力的支持.

通过将因果方法引入解耦表征学习, 可以明确识别并解耦那些影响观测数据的因果因子. 因果解耦表征学习特别适用于那些存在多个生成因子且这些因子之间存在潜在因果关系的场景. 但在实践中, 准确地指定潜在的因果结构并捕捉数据中所有的因果关系可能是困难的. 为了解决这些挑战, 该领域采用了多种技术和算法, 如因果推断、因果发现和结构方程建模.

3 因果解耦表征学习

因果解耦表征学习这一新兴研究领域的核心任务是挖掘数据中蕴含的因果联系, 并得到能够体现这些联系的解耦表示. 首先, 因果关系本身极为复杂, 尤其在高维数据环境中, 如何准确识别并分离不同的因果因素成为一项艰巨任务. 此外, 尽管因果图模型在描述因果结构方面有一定优势, 但在处理复杂动态、反馈机制或未观测变量时, 其能力受到限制, 难以全面捕捉所有因果关系. 在因果推断领域, 研究者们通过融合因果推断、因果建模等方法进行探索实现高效解耦. 下面从融合结构因果模型和基于流模型两个角度对因果解耦表征学习方法进行归纳总结. 结构因果模型通过 DAG 和结构方程模拟变量之间的直接因果效应, 并用于预测干预和反事实的结果. 这种方法擅长揭示静态因果关系, 是因果推断领域的重要工具. 相比之下, 基于流模型的方法则侧重于因果效应在系统中的传播路径. 通过结合正则化流或自回归流, 该模型能够捕捉因果因素之间的复杂交互, 特别是在动

态场景下的因果关系建模. 两类分类的工作总结如表 1 所示. 这些方法能够更准确地捕捉和理解数据背后的复杂因果关系, 为后续提升因果解耦表征学习性能提供参考和借鉴.

表 1 因果解耦表征学习分类

分类	工作	描述	适用范围
融合结构因果模型的解耦表征学习	CausalGAN[16]	明确引入了因果图, 将生成过程建模为因果隐含生成模型, 能够从观测和干预分布中采样	适用于具有明确因果关系且带有标签的数据
	Suter等人[17]	提出的解耦指标IRS是基于结构因果模型框架来定义和计算的	指标适用于量化深度潜在变量模型的鲁棒性
	CausalVAE[18]	引入因果层将独立的外源因素转化为因果内源因素, 从而学习数据中因果相关的概念	特征之间存在明确的因果结构
	Brehmer等人[19]	提出隐式潜在因果模型, 通过神经解码函数隐含地表示因果结构和变量	适用于需要从未标记的低层次数据中学习高层次因果表示的场景
	SCM-VAE[20]	假设一个非线性的结构因果模型来学习可识别的因果表示	具有先验因果结构知识, 非线性因果关系的数据
	CDG-VAE[21]	引入结构因果模型扩展解耦表示的建模	具有因果关系的数据
	Reddy等人[22]	因果生成过程的角度研究解耦表征, 考虑生成因素之间可能存在混杂因素; 并提出因果解耦过程的定义	指标适用于存在潜在混杂因素的场景
	DEAR[23]	利用结构因果模型作为生成模型的先验	具有明确先验因果结构的数据
	CausalDiffAE[24]	基于扩散概率模型和融合结构因果模型	具有复杂语义结构的数据
基于流的解耦表征学习	ICM-VAE[25]	提出结构因果流层, 采用非线性且可学习的流模型参数化因果机制	简单数据集的干预场景
	CauF-VAE[26]	在自回归流中引入因果条件器和邻接矩阵实现因果流	具有复杂因果关系

3.1 融合结构因果模型的解耦表征学习

机器学习方法的性能在很大程度上取决于它们应用的数据表示 (或特征). 人工智能的飞跃进步要通过学习识别和解开隐藏在低级感官数据观察中的潜在解释因素来实现. 为了更好地解耦学习, 可以从因果推理的角度来评估模型与底层因果结构的匹配程度, 学习未知因果混杂变量的分解和结构化因果表征[27]. 具体对比内容如表 2 所示.

表 2 融合结构因果模型的因果解耦表征学习方法对比

工作	优点	缺点	评价指标	数据集
CausalGAN[16]	结合了因果图和生成对抗网络, 能够从观测和干预分布中采样	在训练过程中, 生成器可能会陷入标签条件下的模式崩溃	—	CelebA
Suter等人[17]	IRS指标为评估深度表示的鲁棒性提供了一个实用的工具; 提出解耦因果过程的定义, 为后续研究提供了理论基础	对数据集要求较高	IRS, MI, INFO, MI	dSprites
CausalVAE[18]	能够生成具有因果语义的反事实数据, 支持干预操作	需要预先定义因果结构, 在实际应用中可能难以实现, 特别是在因果关系复杂的场景中	MIC, TIC	Pendulum, Flow, CelebA
Brehmer等人[19]	证明在弱监督的设置下, SCM (包括因果变量和结构)是可被识别的, 为在有限的监督信息下学习复杂的因果结构提供了理论基础	隐式潜在因果模型的表示可能不如显式模型直观, 解释性稍差	DCI, SHD, 干预准确性	Causal3DIdent, CausalCircuit
SCM-VAE[20]	非线性的结构因果模型结构, 能够更好地捕捉数据中的复杂因果关系	依赖于非线性结构因果模型的先验假设	MIC, TIC	Pendulum, Flow, CelebA
CDG-VAE[21]	能够实现因果解耦表示和因果解耦生成	比较依赖DAG结构	样本效率, 分布鲁棒性, CDM	Pendulum, Tabular
Reddy等人[22]	提出了新的评价指标, 拓展研究	模型的性能高度依赖于潜变量的选择	IRS, DCI, UC, CG	CANDLE, dSprites, MPI3D-Toy
DEAR[23]	能够在有限的监督信息下学习复杂的因果结构	联合训练生成器、编码器和SCM参数需要大量的计算资源, 并且训练过程不稳定	样本效率, 分布鲁棒性	Pendulum, CelebA
CausalDiffAE[24]	能够生成高质量的图像, 保证生成过程的可控性	扩散模型的性能直接影响CausalDiff-AE的生成效果	DCI, Effectiveness	MorphoMNIST, Pendulum, CausalCircuit

最早期的工作追溯到 Kocaoglu 等人[16]提出 CausalGAN 是因果隐含生成模型, 能够从给定的因果图中生成观测和干预分布. 同时 CausalGAN 强调因果图在生成模型中的重要性, 展示了如何通过结构化生成器架构来反映因果图的结构, 从而训练出能够进行观测和干预采样的生成模型.

为拓展解耦表征学习的研究, Suter 等人[17]提出了一种评价指标, 利用干预鲁棒性分数 (interventional robustness score, IRS) 来量化评估深度潜在变量模型在面对干预时的表现, 并且利用结构因果模型来研究特征表示对干预的效应. 虽然没有实现真正意义上的解耦, 但是为因果解耦表征学习提供了新的度量指标, 助力因果解耦表征学习的发展.

随着深入的研究, 第 1 个真正意义上实现因果解耦的工作是 Yang 等人[18]提出的 CausalVAE 框架, 为因果解耦表征学习带来新的视角与见解. CausalVAE 从因果关系的角度考虑数据中变异因素之间的关系, 通过利用结构因果模型的原理, 实现了对数据中因果关系的解耦表示学习. CausalVAE 中的因果层, 本质上采用的就是结构因果模型的原理, 将独立的外生因素转化为与数据中的因果相关概念相匹配的内生因素, 使用掩码机制将父变量传递到子变量上, 模拟 SCM 的赋值操作, 同时这样的因果层支持干预操作, 执行“do-operation”操作, 从而生成具有因果语义的反事实数据. 图 1 是 CausalVAE 总体模型结构图, 其中的①、②分别是推断过程和生成过程. 其中推理过程, 模型通过编码器学习到独立噪声 (ε_1, ε_2, ε_3, ε_4) 经过因果层生成富含语义信息的潜在表示 (z_1, z_2, z_3, z_4); 生成过程, 将潜在表示以及附加信息 u 被用来重建图像, 从而实现对原始数据的准确还原, 具体操作是图中的③, 将获得的因果表示 z 经过掩码层操重构自身. 掩码层使用一个邻接矩阵 A, 表示各节点之间的关系, 允许模型在给定父节点变量的情况下重建子节点变量, 这是实现对因果系统进行干预的关键. 在干预操作中, 通过将 z_i 的值固定为一个特定的值, 可以模拟对因果系统中的某个部分进行外部控制的效果如图 1 中④所示.

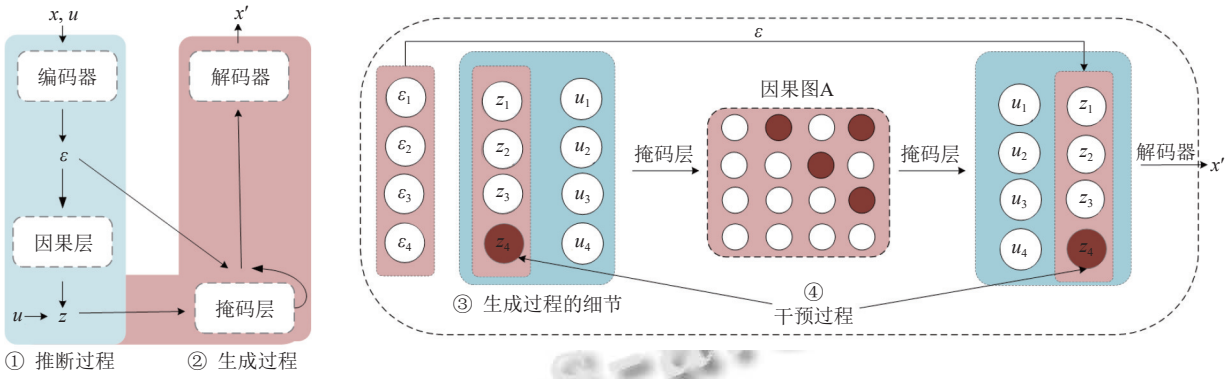

图 1 CausalVAE 总体模型结构图

与 CausalVAE 的有监督方案不同, Brehmer 等人[19]提出了一种基于变分自编码器的弱监督框架, 引入了隐式潜在因果模型, 即无需优化显式离散图结构即可表示因果变量和因果结构, 并证明在弱监督的设置下, 即仅有观测数据对 (干预前后) 而缺乏具体干预标签的情况下, SCM (包括因果变量和结构) 是可被识别的. 这一发现为在有限的监督信息下学习复杂的因果结构提供了理论基础. Komanduri 等人[20]提出 SCM-VAE 框架在 CausalVAE 的基础上假设一个非线性的结构因果模型结构, 提出结构因果先验, 它通过正则化后验并强制执行与因果图一致的潜在表示维度之间的因果结构. 相比之下, CausalVAE 中的条件因子先验简单地假设因子之间的相互独立性.

同时为进一步探索因果解耦表征学习, Reddy 等人[22]进一步分析数据之间的关系提出了生成式潜变量模型 (如 VAE) 实现因果解耦应满足的基本属性. 这些基本属性都与 SCM 密切相关, 包括变量的独立性、潜在变量与生成因素之间的一一对应关系, 以及生成因素对输出的直接因果效应. 这些属性确保了潜变量模型在捕捉和模拟数据的因果结构方面具有精确性, 使得潜变量模型不仅停留在数据的表象层面, 而是能够触及数据生成的根本原因. 同时提出两个新的评估指标——无混杂性 (unconfoundedness, *UC*) 和反事实生成性 (counterfactual generativeness, *CG*). 这两

个指标用于评估潜在变量模型在因果解耦方面的表现.

为提升解耦性能, 将解耦与稀疏相结合. Lachapelle 等人[28]提出了一种表征学习方法, 通过同时学习潜在因素和稀疏因果图模型来诱导解耦. 这种方法基于机制稀疏性原则, 假设潜在因素之间的依赖关系是稀疏的, 并通过正则化技术来诱导解耦. 相比于其他研究工作, Shen 等人[23]提出的 DEAR 是 VAE-GAN 的框架, 这是一种在弱监督信息下进行的生成因果表示的方法. DEAR 的核心思想是利用结构因果模型作为生成模型的先验分布, 这与传统的生成模型中常用的独立同分布的潜在变量先验有所不同. 在 DEAR 框架下, 生成器和编码器不是独立训练的, 而是与 SCM 的参数一起, 通过一个适当的生成对抗网络 (generative adversarial network, GAN) 算法联合训练. 这种训练方式允许模型在生成过程中考虑到潜在变量之间的因果结构, 从而生成的数据不仅在分布上与真实数据相似, 而且在潜在变量的因果关系上也保持一致.

2023 年, An 等人[21]通过整合结构因果模型提出 CDG-VAE 框架, 利用因果图来指导变分自编码器的学习过程, 以实现因果解耦表示和因果解耦生成, 提升下游任务样本效率和分布鲁棒性, 扩展模型的应用范围.

随着扩散模型的广泛应用, Komanduri 等人[24]提出了 CausalDiffAE 框架, 以期碰撞出不一样的火花. CausalDiffAE 融合了结构因果模型和扩散概率模型. 通过学习与因果变量相关的潜在表示, CausalDiffAE 能够实现对生成图像的精确控制, 允许用户通过干预特定的因果变量来生成具有特定属性的图像, 保证生成过程的高质量和可控性.

3.2 基于流模型的解耦表征学习

通过整合先进的流模型和因果推断技术, 可以构建一个更加强大和灵活的系统, 以学习和应用数据中的复杂因果结构.

Monti 等人[29]提出的基于自回归流模型的方法, 通过利用流模型的归一化对数密度估计, 导出了基于似然比的双变量因果方向度量. 这种方法的核心在于, 流模型能够提供一个连续的概率分布, 能够通过比较不同方向的似然比来推断因果关系.

2022 年, Ren 等人[30]使用基于流的潜在变量模型进行因果效应推断. 通过流模型的表达能力和可逆性, 恢复数据中固有的混杂结构, 提高因果效应推断的准确性. 同时, Ren 等人[31]还提出了一种基于流的扰动方法, 用于解决双变量因果发现问题. 利用流模型的编码-解码过程来估计扰动误差, 从而推断因果方向. 上述基于流模型的各种实验为因果解耦表征学习提供了新的视角和方法.

基于流的方法在因果推断领域已经展现出了显著的优势, 为了充分发挥这一方法的潜力, 可以将其引入解耦表征学习的研究中, 从而为该领域带来全新的视角和丰富的研究内容. Komanduri 等人[25]提出 ICM-VAE 的框架, 用于学习因果分离的表示. 该框架中的结构因果流层 (structural causal flow, SCF), SCF 层采用的流模型, 通过将变量依次生成, 使每个因果变量只依赖于先前生成的因果变量的子集 (即其父变量). 图 2 中具体描述了结构因果流层将独立的噪声变量 $\varepsilon=(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ 转换为具有因果关系的潜在变量 $z=(z_1, z_2, z_3, z_4)$ 的过程. 具体来说, 通过一系列微分同胚变换 f_i 来实现, 每个变换 f_i, 负责从噪声变量 ϵ_i 和父变量 z_{pa_i} 计算出一个因果变量 z_i, 公式为 $z_i=f_i(\varepsilon_i ; z_{pa_i})$. 其中, 采用的流模型基于仿射自回归流来参数化因果机制, 这些机制负责将噪声变量映射到潜在的因果变量, 进而揭示数据中固有的因果结构. 通过这种方式, ICM-VAE 框架能够对数据中的因果关系进行建模, 而不单纯依赖于提高模型的性能.

为进一步提升在复杂场景的应用, Fan 等人[26]提出的 CauF-VAE 框架通过因果流[32]的机制整合数据中的因果结构信息, 使得模型不仅能够学习数据的潜在表示, 还能够自动捕捉和表达数据中的因果关系. 该因果流的设计是在自回归流模型的基础上, 通过引入邻接矩阵和因果条件器来实现因果流的机制. 其中, 邻接矩阵用于表示变量之间的因果依赖关系, 而因果条件器则在数据转换过程中动态调整模型对因果依赖的考虑, 使得模型能够在生成表示的过程中考虑到变量之间的因果交互作用. 在模型的训练过程中, 通过因果条件器的作用, 确保潜在因子之间的因果关系在生成表示时被正确地传播.

这一方法不依赖于传统的结构因果模型, 即无需依赖预先设定的因果图或因果规则来定义变量间的因果关系. 而是直接通过数据驱动的方式, 在流模型中自适应地学习数据的因果结构特性. 这使得模型能够更加灵活地从数据中捕捉因果关系, 避免了手工指定因果结构的局限性.

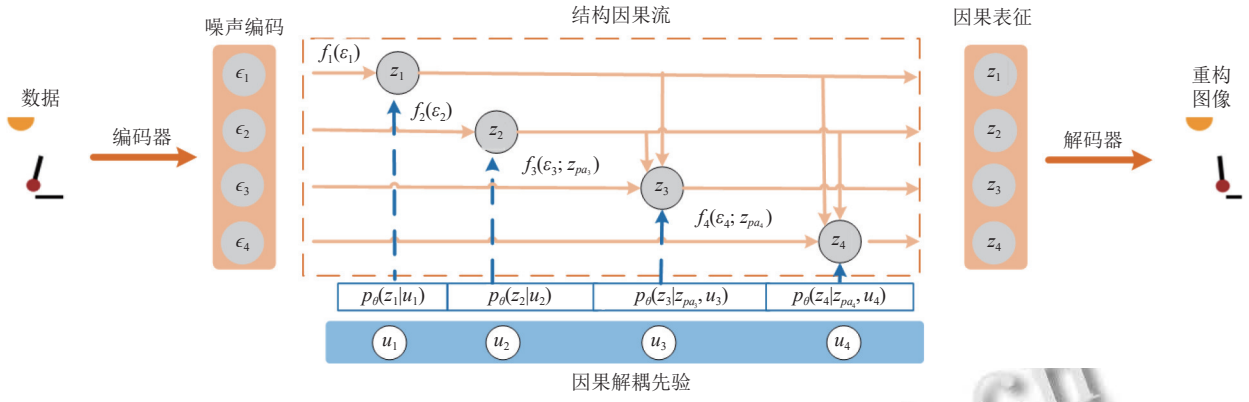

图 2 ICM-VAE 总体模型结构图

3.3 因果解耦表征学习相关的评价指标及数据集

为了评价解耦效果的好坏, 需要相应的数据集和客观的评价指标进行评估. 数据集如表 3 所示, 接下来主要介绍相关的评价指标.

表 3 常用数据集

数据集	类型	描述	复杂性
Pendulum 数据集[18]	合成数据集	每个图像由4个连续的相位生成: 摆角、光线角度、阴影长度和阴影位置	由于其合成特性和明确的潜在因果因子, 模型在该数据集上能较容易地学到因果关系
Flow 数据集[18]	合成数据集	模拟水流动的因果关系, 每张图片包含3个实体 (球、水和孔洞)和4个概念 (球的大小、水面的高度、孔洞的高度和水流速)	
CausalCircuit 数据集[19]	合成数据集	由 4 个真实的潜在因果变量生成. 分别是机器人臂的位置、红色光强度、绿色光强度和蓝色光强度	虽然是合成数据集, 但其涉及多个较为复杂因果变量, 因此需要模型具有较高的表达能力来捕捉潜在的因果结构
CelebA 数据集[33]	真实数据集	由200 000张名人面孔图像组成, 具有40个离散的属性, 每个属性的值都为−1或1	复杂的高维离散数据, 需要更加复杂的解耦方法来处理数据中的潜在因果因素, 特别是在面对多个相关属性和可能的干扰时

互信息差 (*MIG*) 是一种新的信息论度量方法来评估解耦表示的质量. *MIG* 是一种无分类器的度量方法, 可以推广到任意分布的和非标量潜在变量. *MIG* 通过计算每个真实生成因素 v_k 与所有潜在变量 z_j 之间的归一化互信息, 然后取最大的互信息与次大的互信息之间的差值, 如式 (4) 所示. *MIG* 的值越接近 1, 表示解耦程度越高, 潜在变量与真实生成因素之间的对应关系越清晰. *MIG* 指标衡量每个因素在解耦表示中最高和第 2 高的坐标之间的互信息差距. 当因素之间存在相关性时, *MIG* 指标可能无法准确评估解耦效果.

$$MIG = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{H(v_k)}\left(I_n(z_{j^{(k)}};v_k)-\max_{j\neq j^{(k)}}I_n(z_j;v_k)\right) \quad (4)$$

SAP (separated attribute predictability) 由 Kumar 等人[34]提出的一种衡量解耦表示性能的方法. 首先, 构建一个 $d\times k$ 的得分矩阵, 其中 d 是潜在变量的数量, k 是生成因素的数量. 矩阵中的第 i, j 个条目 S_{ij} 是使用仅第 i 个潜在变量来预测第 j 个生成因素的线性回归或分类得分. 对于得分矩阵 S 的每一列, 计算前两个得分条目的差异, 然后取这些差异的平均值作为最终的 SAP 分数. SAP 分数越高表明生成因素主要由单一潜在维度捕获, 反之, 则表明潜在表示可能未能很好地解耦生成因素.

DCI 由 Eastwood 等人[35]定义和量化 3 个标准来评估解耦表示的质量, 包括解耦度、完整性和信息量. 其中, 解耦度衡量每个变量捕获的生成因素的数量; 完整性衡量每个生成因素被单个代码变量捕获的程度, 理想情况下, 每个生成因素由一个代码变量完全捕获; 信息量衡量表示关于生成因素的变化所包含的信息量. 同时该指标仅适用于生成因子之间相互独立的情况.

最大信息系数 (maximal information coefficient, *MIC*)[36]和总信息系数 (total information coefficient, TIC) 是在解耦表征学习中计算学习到的表示与真实标签之间的信息相关度. *MIC* 衡量两个变量之间依赖性的统计量, 它通过最大化互信息的估计值来确定变量间的关系强度. *MIC* 值越高, 表明学习到的表示与真实

标签之间的依赖性越强. 但是 *MIC* 对某些噪声不敏感, 并且由于 *MIC* 要求测试每个数据集的所有可能的分箱方案, 这使得 *MIC* 的计算复杂度很高. 具体如式 (5) 所示. TIC 衡量两个变量之间总体信息关系的统计量, 通过等特征矩阵的所有条目总和来计算. TIC 值越高, 表明学习到的表示与真实标签之间的信息关系越紧密.

$$MIC = \max_{|X||Y|<B} \frac{I[X;Y]}{\log_2(\min(|X|,|Y|))} \tag{5}$$

UC 指标用于评估不同生成因素是否被独特的潜在维度所捕捉, 且没有重叠. 如果一个模型能够将每个生成因素映射到一个独特的潜在维度, 且这些映射之间没有交集, 那么称这个潜在空间是无混杂的. *UC* 指标通过计算所有可能的潜在变量对之间的 Jaccard 相似性系数来衡量, 如式 (6) 所示. *CG* 指标用于评估在潜在空间中实现无混杂性后, 模型是否能够灵活且受控地生成反事实实例, 具体如式 (7) 所示. 但 *UC* 和 *GC* 通常假设生成因素是条件独立的.

$$UC = 1 - E_{x\sim p_X}\left[\frac{1}{S}\sum_{I,J}\frac{|Z_I^x \cap Z_J^x|}{|Z_I^x \cup Z_J^x|}\right] \tag{6}$$

$$CG = E_I\left[\left|ACE_{Z_I^X}^{X_I^{cf}} - ACE_{Z_{\setminus I}^X}^{X_{\setminus I}^{cf}}\right|\right] \tag{7}$$

样本效率和分布鲁棒性.样本效率通常用来衡量模型在有限样本下的学习性能, 即以 100 个样本的测试准确率与全部样本的测试准确率之比来衡量. 这个比例反映了模型在样本数量受限时的表现与在拥有全部数据时的表现之间的差异. 一个模型如果样本效率较高, 那么它能够在数据较少的情况下达到较高的准确率, 这在数据获取成本较高或数据较为稀缺的场景中尤为重要. 实际应用中, 通常难以获得大量标注数据, 样本效率能够衡量模型在少量数据上学习有效解耦表征的能力.

分布鲁棒性衡量的是模型在面对数据分布变化时的可靠性和稳定性, 具体的方法是首先人为地在训练数据上通过注入虚假相关性引入分布偏移, 然后在没有分布偏移的测试数据上评估模型性能, 记录平均测试准确率和最差情况测试准确率. 同时, 现实世界中生成因素之间通常存在复杂的因果关系, 这意味着数据分布可能会发生变化. 好的表征应该能够在不同的数据分布下保持一致性和稳定性, 因此用分布鲁棒性评估模型有利于全方位评估模型的性能.

4 因果解耦表征学习的应用

4.1 因果解耦表征学习在图像生成中的应用

在图像生成领域, 解耦表征学习技术能够识别并学习目标图像中的独立因素, 从而揭示与潜在表示一致的关键因子, 并更精确地控制图像生成过程. 在传统的图像生成方法中, 生成模型通常依赖于概率模型来学习数据的分布, 并生成符合该分布的样本. 最典型的框架是生成对抗网络和变分自编码器. 这些方法通过直接优化数据生成过程来捕捉数据的统计特性. FactorVAE[37]通过增加边际分布与其实各项边际分布乘积之间的 KL 散度, 直接鼓励每个维度的表示是独立的, 从而可以在图像生成中对每个维度进行操作, 生成多样化的图片. JointVAE[38]在无监督的情况下学习解耦的、可解释的联合连续和离散表示, 它特别关注于从数据中自动发现连续和分类的变化因素, 通过学习连续和离散的潜在变量, 模型能够适应不同的图像生成任务. Wang 等人[39]提出 DisCo 框架, DisCo 利用变分自编码器 (VAE) 作为背景编码器, 将背景图像转换为密集特征图, 以保留复杂的背景细节. 同时, 使用一个小的卷积编码器处理高度抽象的骨架, 以控制姿势. 对于人类主体, DisCo 结合其 CLIP 图像嵌入与去噪 U-Net, 通过交叉注意力模块帮助动态前景合成.这种解耦控制使得模型能够灵活地组合不同的人类主体、背景和姿势, 生成高质量的图像. 但是, DisCo 难以应用于多人场景和人与物体的交互的场景.

现在越来越多的研究将因果关系纳入解耦表征学习中, 例如 CausalVAE 引入了因果结构和弱监督, 支持生成具有因果语义的图像并创建反事实结果. 意味着 CausalVAE 不仅能够生成看起来真实的图像, 还能够模拟在特定干预下图像可能发生的变化, 这对于理解图像中的因素如何影响整体结构和内容具有重要意义. DEAR 通过 VAE-GAN 架构将因果关系作为先验补充, 利用因果信息来指导生成过程, 能够生成与特定因素变化一致的图像. 这种方法的优势在于, 它能够生成不仅在视觉上逼真, 而且在因果关系上与特定干预一致的图像, 这对于模拟和预测特定变化的影响非常有用.

利用因果解耦表征学习可以生成更多加入干预的图片, 这种干预可以是微小的变化, 也可以是根本性的改变. 这种能力使得因果解耦表征学习在图像生成领域具有巨大的潜力, 因为它不仅能够提高生成样本的

多样性, 还能够提供对生成过程的精细控制. 例如, 在医学成像中, 这种技术可以用来模拟不同治疗方案对患者的影响; 在娱乐产业, 可以用来创建具有特定特征的角色或场景; 在科学研究中, 可以用来模拟不同条件下的实验结果. 总的来说, 解耦表征学习通过引入因果关系, 为图像生成提供了一种新的、强大的工具, 它不仅能够提高生成图像的质量, 还能够增强我们对数据背后因果机制的理解.

4.2 因果解耦表征学习在 3D 姿态中的应用

人体姿态估计, 涉及从图像和视频等输入数据中识别人体部位并构建身体骨架, 已被广泛应用于人机交互、运动分析、增强现实和虚拟现实等领域. 主要分为 2D 人体姿态估计领域和 3D 人体姿态估计领域两大类, 这两类又细分为单人姿态估计和多人姿态估计.其中在 2D 人体姿态估计领域, 单人姿态估计主要采用基于回归的方法和基于热图的方法; 而多人姿态估计则主要采用自上而下的方法和自下而上的方法. 在 3D 人体姿态估计领域, 单人姿态估计方面主要是由 2D 图像推断 3D 人体姿态; 多人姿态估计方法同样分为自上而下和自下而上两大类. 自上而下方法与单人 3D 姿态估计类似, 但需要额外处理多人场景中的视角和遮挡问题; 而自下而上方法则是先检测所有身体关节和深度图, 然后根据根深度和部分相对深度将身体部位关联到每个人[40,41].

深度学习模型倾向于学习到虚假性相关性, 在泛化到训练数据集之外的场景时面临挑战, 尤其是在面对不熟悉的子领域或新的视角时 (即跨域姿态估计), 训练良好的模型往往难以准确识别关节位置. 这种局限性可能源于数据集的偏差或所谓的捷径学习, 导致深度学习模型容易学习到基于统计关联的数据集特定虚假关联[42]. 当这些关联在不同领域间不一致时, 问题就变得尤为突出.

针对上述问题, Zhang 等人[43]提出了一种新颖的跨域姿态估计方法, 该方法结合了因果表示学习和生成对抗网络, 通过生成反事实特征来模拟不同域对图像的影响, 帮助模型学习在不同域之间可转移的因果关系. 它促使模型能够从源域 (已观察到的分布) 泛化到目标域 (反事实的分布), 即让模型学会识别和利用从观察数据到反事实数据的底层因果结构. 其中学习的因果表示主要包括两个分支观察表示分支包括特征提取器 f, 它从源域的图像输入中提取表示, 并在观察到的分布上做预测. 反事实表示分支由一个特征生成器 g 组成, 它从真实的姿态和随机噪声中产生反事实特征. 最小化观察表示的和反事实表示之间的差异距离, 鼓励模型学习从观察分布泛化到反事实分布的底层不变性.

4.3 因果解耦表征学习在无监督领域中的应用

随着大数据和深度学习技术的快速发展, 人工智能模型在许多领域取得了显著的成功. 然而, 传统的监督学习方法往往依赖于大量的标注数据, 这在实际应用中常常受到限制. 尤其是在目标领域缺乏标注数据的情况下, 无监督领域自适应 (unsupervised domain adaptation, UDA), 通过源领域的标注数据与目标领域的无标注数据之间的迁移, 减少由于领域间数据分布差异带来的负面影响.

然而, 现有的 UDA 方法仍然存在着诸多局限性. 首先, 大多数方法过于关注特征层次上的对齐, 忽略了源领域和目标领域之间的潜在因果关系. 其次, 传统的 UDA 方法大多假设源领域和目标领域之间的分布差异仅是简单的统计差异, 但实际上, 源领域和目标领域之间可能存在复杂的结构性差异, 这些差异不仅体现在视觉、语音等表象特征上, 还可能涉及深层次的因果因素, 导致直接的特征对齐方法无法有效捕捉到任务相关的知识. 现有的大多数 UDA 方法通常依赖于对抗训练等技术来进行领域对齐, 但这些方法在训练过程中往往面临着不稳定的问题, 尤其是在没有足够标注数据的情况下, 模型容易陷入局部最优解, 导致性能波动较大.

为了无监督域适应中的域偏差问题, Wang 等人[44]提出一种基于因果解耦的无监督领域自适应方法. 通过引入因果机制, 将语义特征分为因果特征和非因果特征. 因果特征捕捉了输入数据与标签之间的深层次因果关系, 而非因果特征则是由数据偏差或表面统计关系引起的噪声. 通过解耦源领域和目标领域的语义信息, 模型能够在不依赖标签配对数据的情况下, 进行更加准确的知识迁移.

未来的工作可以将特征解耦技术和因果机制应用于更具挑战性的跨媒体分析任务, 以提高模型在新领域的适应性和泛化能力, 进一步探索如何在更多实际应用场景中推广这种方法, 特别是在多模态数据和复杂任务中的应用.

5 因果解耦表征学习未来的研究方向

因果解耦表征学习旨在揭示和分离数据中由潜在因果机制驱动的因子与那些非因果但统计相关的因子, 核心优势在于通过有效的因果推理, 从因果关系的角度进行建模, 提升模型的可解释性、稳定性和泛化能力. 然而, 当前因果解耦表征学习仍面临诸多挑战, 未来的研究方向将集中在理论框架、可解释性与公平性以及评价指标 3 个主要方面.

5.1 因果推理与解耦表征的结合: 构建统一的理论框架

尽管解耦表征学习逐渐受到越来越多的追捧, 但缺少一个普遍认可的定义, 且理论基础尚需加强. 目前, 大部分的因果解耦方法仍然依赖经验性设计, 缺少系统的因果推理框架来明确建模因果关系. 要推动因果解耦表征学习的发展, 首先需要在理论上进一步明确因果解耦的目标和实现机制, 并构建一个统一的因果推理框架.

因果推理框架的设计是未来因果解耦表征学习研究的关键. 传统的解耦表征学习侧重于特征的独立性, 忽略了数据背后的因果关系. 为了解决这个问题, 未来的研究应该将因果学习方法和深度学习方法融合, 从因果推理、因果建模的角度设计解耦目标函数, 明确区分因果因素和非因果因素. 未来的研究应关注如何在学习过程中有效地识别和利用因果关系, 以便在不同任务和环境中实现更强的泛化能力和鲁棒性.

同时, 考虑到实际应用中因果关系的复杂性和未知性, 无监督因果结构学习将成为研究的新焦点, 通过无监督或弱监督学习从数据中自动发现因果关系, 为因果解耦提供理论基础, 并拓宽其应用范围.

5.2 可解释性与公平性: 提升模型透明度与减少偏见

因果解耦表征学习的一个显著优势是增强模型的可解释性. 通过因果推理, 解耦后的表征能够明确区分哪些特征是由因果关系驱动的, 哪些特征只是表面相关, 从而使得模型的决策过程变得更加透明. 在医疗、金融等高风险领域, 能够理解模型的决策逻辑至关重要. 通过引入因果解耦, 研究者可以揭示模型如何根据因果因素进行预测, 帮助专家理解模型的推理过程.

李雅婷等人[45]指出解耦表征能够在抽象推理任务中提供较高的可解释性, 因此将解耦表征引入到实际应用中, 可以帮助系统更好地解释每个预测背后的因果机制. 未来的研究应注重如何将因果解耦与领域专家知识相结合, 以提升模型的可解释性和信任度. 例如, 在医学影像分析中, 因果解耦能够帮助识别哪些图像特征与疾病之间存在因果关系, 从而为医生提供更为透明的决策支持.

同时, 公平性也是因果解耦表征学习的关键议题. 随着机器学习模型的广泛应用, 模型偏见问题日益凸显. 因果解耦表征有助于识别并消除模型中的不公平因素, 如性别、种族或年龄等无关因素, 防止模型对某些群体产生偏见预测. 因此, 未来研究需探索如何在因果解耦学习框架下结合公平性约束, 确保模型在解耦过程中不仅性能优越, 而且能在不同群体间避免不公平预测. 特别是在司法、招聘、贷款审批等敏感领域, 公平性尤为重要, 这要求模型透明度的提升与偏见的减少并行不悖, 共同构建一个更加公正可靠的 AI 系统.

5.3 完善评价指标

当前解耦表征学习领域尚未形成一套统一的度量指标体系, 这一缺陷可能导致理论研究与实际应用之间的断裂. 尽管已有的度量指标, 例如 *MIG*、IRS 和 *MIC*, 在评估潜在因子独立性方面发挥了作用, 但它们在衡量因果关系的复杂性和动态性方面各有局限[46]. 这些指标的不足在于未能全面捕捉因果结构的深度和动态变化, 从而限制了模型的泛化能力. 因此, 未来的研究方向应聚焦于融合因果图和因果推理的方法, 以构建新的解耦表征评估指标. 这些新指标不仅要评估因子间的独立性, 更要深入探讨因果结构的可解释性, 揭示潜在因子间的因果联系和相互作用. 通过这种多维度的评估框架, 可以更准确地理解和优化模型的泛化能力和实际应用效果, 为解耦表征学习模型在多样化应用场景下的表现提供更全面和精细的衡量, 进而促进理论研究与实际应用之间的有效衔接. 这样的综合评估指标将有助于弥补现有度量体系的缺陷, 确保模型在实际应用中的有效性和可靠性.

6 结语

随着人工智能技术, 尤其是机器学习和深度学习领域的快速发展, AI 系统在图像识别、自然语言处理和复杂决策支持系统等任务中展现出卓越能力. 在这一进程中, 解耦表征学习经历了从独立成分分析到变分自编码器的演化, 并逐步引入因果推断和因果建模, 形成了如今的因果解耦表征学习. 这一领域的研究不断深化, 强调了识别和建模变量间因果关系的重要性.

本文综述了因果解耦表征学习的最新研究进展, 涵盖了从理论基础到实际应用的各个方面, 展现了该领域的深度与广度. 文章探讨了如何结合结构因果模型与基于流模型的方法来实现有效的因果解耦, 并讨论了因果解耦表征学习在图像生成、3D 姿态估计和无监督领域适应等应用中的潜力. 同时, 对未来的研究方向进行了展望.

在未来, 因果解耦表征学习应该在理论和应用两个层面上进一步发展. 一方面, 研究者们应该探索更加精细的理论框架, 以便更好地理解和建模数据中的因果结构; 另一方面, 随着研究的不断深入, 因果解耦表征学习应该在解释性、泛化能力和决策支持等方面为人工智能的发展提供新的动力, 促进其在多个领域取得重要进展.

参考文献

1 Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798–1828. [doi: 10.1109/TPAMI.2013.50]

2 文载道, 王佳蕊, 王小旭, 等. 解耦表征学习综述. 自动化学报, 2022, 48(2): 351–374. [doi: 10.16383/j.aas.c210096]

3 Schölkopf B, Locatello F, Bauer S, *et al*. Toward causal representation learning. Proceedings of the IEEE, 2021, 109(5): 612–634. [doi: 10.1109/JPROC.2021.3058954]

4 Locatello F, Bauer S, Lucic M, *et al*. Challenging common assumptions in the unsupervised learning of disentangled representations. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 4114–4124.

5 Sikka H. A deeper look at the unsupervised learning of disentangled representations in β-VAE from the perspective of core object recognition. arXiv:2005.07114, 2020.

6 Kingma DP, Welling M. Auto-encoding variational Bayes. Proceedings of the 2nd International Conference on Learning Representations. Banff, 2014.

7 Higgins I, Matthey L, Pal A, *et al*. β-VAE: Learning basic visual concepts with a constrained variational framework. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017. 3.

8 Chen RTQ, Li XC, Grosse R, *et al*. Isolating sources of disentanglement in VAEs. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 2615–2625.

9 Locatello F, Poole B, Rätsch G, *et al*. Weakly-supervised disentanglement without compromises. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 6348–6359.

10 Locatello F, Tschannen M, Bauer S, *et al*. Disentangling factors of variation using few labels. arXiv:1905.01258, 2019.

11 Sanchez EH, Serrurier M, Ortner M. Learning disentangled representations via mutual information estimation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 205–221.

12 Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. New York: Basic Books, 2018.

13 Peters J, Janzing D, Schölkopf B. Elements of causal inference: Foundations and learning algorithms. Cambridge: The MIT Press, 2017.

14 Pearl J. Causality. 2nd ed., Cambridge: Cambridge University Press, 2009.

15 Parascandolo G, Kilbertus N, Rojas-Carulla M, *et al*. Learning independent causal mechanisms. Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan: PMLR, 2018. 4033–4041.

16 Kocaoglu M, Snyder C, Dimakis AG, *et al*. CausalGAN: Learning causal implicit generative models with adversarial training. arXiv:1709.02023, 2017.

17 Suter R, Miladinovic D, Schölkopf B, *et al*. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 6056–6065.

18 Yang MY, Liu FR, Chen ZT, *et al*. CausalVAE: Disentangled representation learning via neural structural causal models. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9588–9597.

19 Brehmer J, De Haan P, Lippe P, *et al*. Weakly supervised causal representation learning. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 2776.

20 Komanduri A, Wu YK, Huang W, *et al*. SCM-VAE: Learning identifiable causal representations via structural knowledge. Proceedings of the 2022 IEEE International Conference on Big Data (Big Data). Osaka: IEEE, 2022. 1014–1023.

21 An SH, Song K, Jeon JJ. Causally disentangled generative variational AutoEncoder. Frontiers in Artificial Intelligence and Applications, 2023, 372: 93–100. [doi: 10.3233/FAIA230258]

22 Reddy AG, Godfrey LB, Balasubramanian VN. On causally disentangled representations. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 8089–8097.
23 Shen XW, Liu FR, Dong HZ, *et al*. Weakly supervised disentangled generative causal representation learning. The Journal of Machine Learning Research, 2022, 23(1): 241.
24 Komanduri A, Zhao C, Chen F, *et al*. Causal diffusion Autoencoders: Toward counterfactual generation via diffusion probabilistic models. Amsterdam: IOS Press, 2024. 2516–2523.
25 Komanduri A, Wu YK, Chen F, *et al*. Learning causally disentangled representations via the principle of independent causal mechanisms. arXiv:2306.01213, 2023.
26 Fan D, Kou YN, Gao CH. CauF-VAE: Causal disentangled representation learning with VAE and causal flows. arXiv: 2304.09010, 2023.
27 Wang X, Chen H, Tang SA, *et al*. Disentangled representation learning. arXiv:2211.11695v4, 2024.
28 Lachapelle S, Rodriguez P, Sharma Y, *et al*. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. Proceedings of the 1st Conference on Causal Learning and Reasoning. Eureka: CLeaR, 2022. 428–484.
29 Monti RP, Khemakhem I, Hyvärinen A. Autoregressive flow-based causal discovery and inference. arXiv:2007. 09390, 2020.
30 Ren SG, Li DC, Li P. Causal effect prediction with flow-based inference. Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM). Orlando: IEEE, 2022. 1167–1172.
31 Ren SG, Li P. Flow-based perturbation for cause-effect inference. Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta: ACM, 2022. 1706–1715.
32 Khemakhem I, Monti RP, Leech R, *et al*. Causal autoregressive flows. Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. PMLR, 2021. 3520–3528.
33 Liu ZW, Luo P, Wang XG, *et al*. Deep learning face attributes in the wild. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3730–3738.
34 Kumar A, Sattigeri P, Balakrishnan A. Variational inference of disentangled latent concepts from unlabeled observations. arXiv:1711.00848, 2017.
35 Eastwood C, Williams CKI. A framework for the quantitative evaluation of disentangled representations. Proceedings of the 6th International Conference on Learning Representations. Vancouver: OpenReview.net, 2018.
36 Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(9): 3354–3359.
37 Kim H, Mnih A. Disentangling by factorising. Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan: PMLR, 2018. 2654–2663.
38 Dupont E. Learning disentangled joint continuous and discrete representations. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 708–718.
39 Wang T, Li LJ, Lin K, *et al*. DisCo: Disentangled control for realistic human dance generation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2024. 9326–9336.
40 Zheng C, Wu WH, Chen C, *et al*. Deep learning-based human pose estimation: A survey. ACM Computing Surveys, 2023, 56(1): 11.
41 Munea TL, Jembre YZ, Weldegebriel HT, *et al*. The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. IEEE Access, 2020, 8: 133330–133348. [doi: 10.1109/ACCESS.2020.3010248]
42 Acharya J, Bhattacharyya A, Daskalakis C, *et al*. Learning and testing causal models with interventions. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 9469–9481.
43 Zhang XH, Wong Y, Wu XF, *et al*. Learning causal representation for training cross-domain pose estimator via generative interventions. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 11250–11260.
44 Wang SS, Chen YY, He ZW, *et al*. Disentangled representation learning with causality for unsupervised domain adaptation. Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023. 2918–2926.
45 李雅婷, 肖晶, 廖良, 等. 面向视觉数据处理与分析的解耦表示学习综述. 中国图象图形学报, 2023, 28(4): 903–934. [doi: 10.11834/jig.211261]
46 成科扬, 孟春运, 王文杉, 等. 解耦表征学习研究进展. 计算机应用, 2021, 41(12): 3409–3418. [doi: 10.11772/j.issn.1001-9081.2021060895]

(校对责编: 张重毅)