中国科学:信息科学 在审文章: 1–43

SCIENTIA SINICA Informationis

论文

解耦表征学习: 方法、应用及模型设计

王鑫[1,2], 秦晨阳[1], 朱文武[1,2*]

1. 清华大学计算机科学与技术系, 北京 100084
2. 北京信息科学与技术国家研究中心, 北京 100084
* 通信作者. E-mail: wwzhu@tsinghua.edu.cn

国家重点研发计划项目（批准号：2023YFF1205001）和北京信息科学与技术国家研究中心（批准号：BNR2026TD03005）资助

摘要 解耦表征学习 (Disentangled Representation Learning) 旨在学习一种模型, 使其能够识别并分离观测数据中隐藏的潜在因素, 并以表征形式加以刻画. 将潜在变化因素分解为具有语义意义的变量, 有助于学习可解释的数据表征, 从而模拟人类在观察对象或关系时形成有意义理解的过程. 作为一种通用的学习策略, 解耦表征学习已在多个领域展现出其强大的能力, 能够有效提升模型的可解释性、可控性、鲁棒性以及泛化能力, 广泛应用于计算机视觉、自然语言处理和数据挖掘等场景. 本文从多个角度对解耦表征学习进行了系统的综述, 包括其研究动机、定义、方法论、评估体系、应用场景以及模型设计等方面. 我们首先介绍了两个广为认可的定义, 即直观定义与群论定义, 用于刻画解耦表征学习的核心概念. 随后, 我们从模型类型、表征结构、监督信号以及独立性假设四个维度, 将现有的解耦表征学习方法进行了系统的分类与分析. 此外, 我们探讨了解耦表征学习的应用和在不同场景的模型设计原则, 以便在实际任务中发挥更优的性能. 最后, 本文指出了当前解耦表征学习研究所面临的主要挑战, 并提出了值得未来深入研究的潜在方向. 我们相信, 本研究可为推动解耦表征学习领域的发展与创新提供有益的参考与启示.

关键词 解耦表征学习; 表征学习; 计算机视觉; 模式识别

1 引言

当人类观察一个物体时, 我们通常会基于已有的先验知识, 试图理解该物体的各种属性 (例如形状、大小、颜色等). 然而, 现有的端到端黑箱式深度学习模型往往采用“捷径”策略——直接通过学习物体的表征来拟合数据分布与判别准则 [1], 而未能像人类一样具备提取潜在属性并进行概念化泛化的能力. 为弥补这一差距, 一种重要的表征学习范式——“解耦表征学习”被提出 [2], 并迅速引起了研究界的广泛关注.

解耦表征学习是一种学习范式, 其目标是设计机器学习模型, 使其能够获得能够识别并分离观测数据中隐藏潜在因素的表征. 解耦表征学习有助于学习具有语义意义的、可解释的观测数据表征. 已有研究 [2,3] 表明, 解耦表征学习具备以类似人类的方式理解世界的潜力——这种理解体现在将现实世界观测中的语义信息分解为相互独立的因素. 在表征空间中实现的解耦, 有助于模型学习具有可解释语义的独立因素, 能够在以下三个

引用格式: 王鑫. 秦晨阳. 朱文武. 解耦表征学习: 方法、应用及模型设计. 中国科学: 信息科学, 在审文章
Xin Wang, Chenyang Qin, Wenwu Zhu.
. Sci Sin Inform, for review

方面在多个机器学习任务中展现出显著优势:i) 可解释性: 解耦表征学习能够学习与潜在生成因素一致的、具有明确语义意义且相互独立的表征; ii) 泛化性: 通过将任务相关的表征从原始耦合输入中分离出来, 解耦表征学习能够显著提升模型的泛化能力; iii) 可控性: 通过在潜空间中操作已学习到的解耦表征, 解耦表征学习能够实现可控的生成.

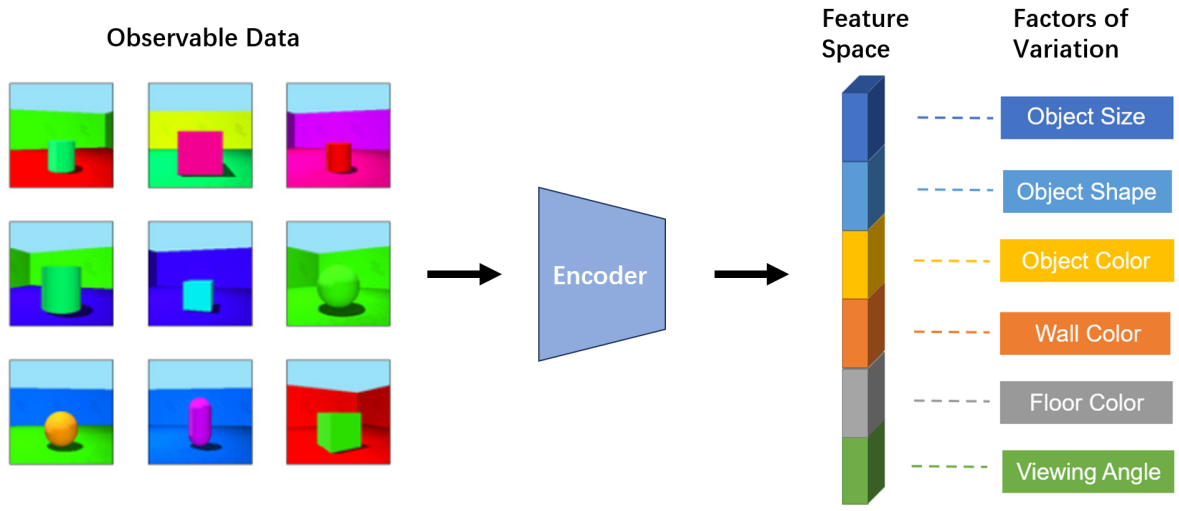

图 1 在 Shape3D [4] 场景中. 右侧六个矩形表示 Shape3D 数据集中的六个相互独立的变化因素. 理想情况下. 解耦表示学习模型应当能够在潜在表示空间中. 使用相互独立的潜变量来编码这些不同的变化因素.

In the Shape3D [4] scene, the six rectangles on the right represent the six independent factors of variation in the Shape3D dataset. Ideally, a disentangled representation learning model is expected to encode these distinct factors of variation using mutually independent latent variables within the latent representation space.

由此, 一个自然的问题出现了: 解耦表征究竟应当学习什么? 答案可追溯至 Bengio 等人 [2] 提出的 “解耦表征” 概念, 其核心在于变化因素. 如图 1 所示, Shape3D 数据集 [4] 是解耦表征学习研究中常用的基准数据集, 包含六种独立的变化因素: 物体大小、物体形状、物体颜色、墙壁颜色、地板颜色以及视角. 解耦表征学习的目标在于将这些因素相互分离, 并将其编码为表征空间中独立且互不干扰的潜变量. 在这种情况下, 控制物体形状的潜变量只会随形状变化而变化, 而在其他因素变化时保持不变. 类似地, 控制其他属性的潜变量亦各自独立, 仅与对应因素的变化相关.

通过理论分析与实证研究, 解耦表征学习在以下三个方面展现出了显著优势:i) 不变性: 解耦表征中的每个元素对于外部语义的变化具有不变性 [5~8];ii) 完整性: 所有解耦表征均能够分别对应真实语义, 并具备生成已观测、未观测, 甚至反事实样本的能力 [9~12]; iii) 泛化性: 所学习的表征是内在且稳健的, 不会捕获混淆或有偏语义, 因此能够在下游任务中表现出良好的泛化能力 [13~15].

在上述动机与需求的推动下, 研究者围绕解耦表征学习的方法、应用和模型设计开展了大量工作. 当前最典型的解耦表征学习方法多基于生成模型 [6,9,16,17], 这些方法最初在学习视觉图像的可解释表征方面显示出巨大潜力. 此外, 基于因果推理 [14] 和群论 [18] 的方法也在解耦表征学习研究中得到广泛采用. 解耦表征学习架构设计的核心思想在于: 在优化固有任务目标 (如生成或判别目标) 的同时, 鼓励潜在因素学习到相互解耦的表征. 凭借其在捕获可解释性、可控性与鲁棒性表征方面的显著能力, 解耦表征学习已被广泛应用于多个领域, 包括计算机视觉 [8,19~22]、自然语言处理 [23~25]、推荐系统 [26~29] 以及图学习 [29,30] 等, 从而显著提升了多种下游任务的性能.

本文的主要贡献如下: 我们首次从理论、方法、评估、应用及模型设计等多个角度, 对解耦表征学习进行了系统而全面的综述. 具体而言:

- 在第 2 节中, 我们介绍了解耦表征学习的定义;
- 第 3 节对现有解耦表征学习方法进行了全面回顾与分类;
- 第 4 节讨论了解耦表征学习实现中的常用评估指标;
- 第 5 节介绍了解耦表征学习在各类下游任务中的应用;
- 第 6 节基于不同任务需求, 探讨了合适的解耦表征学习模型设计策略;
- 最后, 第 7 节总结了当前解耦表征学习研究中存在的开放性问题与未来发展方向.

与我们研究最为相关的工作是 Liu 等人 [31] 的综述, 但其仅聚焦于图像领域及医学影像应用. 相比之下, 本文从更为通用的视角出发, 对解耦表征学习的定义、分类体系、应用范围及模型设计方案进行了全方位的总结与讨论.

2 解耦表征学习的定义

直观定义. Bengio 等人 [2] 提出了解耦表征学习的一个直观定义:

定义 1 解耦表征应能够区分数据中不同、独立且信息丰富的变化因素. 单个潜变量应当对某一变化因素的变化保持敏感, 同时在其他因素发生变化时保持相对不变.

该定义同时指出, 潜在变量在统计意义上应当相互独立. 遵循这一思路, 早期的解耦表征学习方法可追溯至独立成分分析和主成分分析. 随后, 众多基于深度神经网络的研究工作也延续了这种思想 [5~7, 9, 32~37]. 大多数模型与评估指标均假设生成因子与潜在变量在统计上是相互独立的.

定义 1 在文献中被广泛采用, 并构成了第 3 节中多数解耦表征学习方法的基础.

群论定义. 为给出更严谨的数学描述, Higgins 等人 [18] 从群论的角度定义了解耦表征学习, 这一观点随后被大量工作 [38~41] 所采纳. 我们在此简要回顾该基于群论的定义如下:

定义 2 考虑一个对称群 G, 其作用于空间 W(即生成观测数据的真实变化因素空间), 数据空间为 O, 表征空间为 Z. 假设 G 可以分解为直积 $G = G_1 \times G_2 \times \cdots \times G_n$. 若表征 Z 满足以下条件, 则称其在 G_i 上是离散的:

1. 存在群作用 G 作用于 Z: $G \times Z \to Z$;
2. 存在从 W 到 Z 的映射 $f : W \to Z$, 该映射与 G 对 W 与 Z 的作用等价. 此条件可形式化为: $g \cdot f(w) = f(g \cdot w), \forall g \in G, \forall w \in W$, 如图 2 所示.
3. 群 G 对 Z 的作用被 G 的分解所解耦. 换言之, 存在分解 $Z = Z_1 \times \cdots \times Z_n$ 或者 $Z = Z_1 \oplus \cdots \oplus Z_n$, 其中 Z_i 仅受 G_i 影响, 且对所有 $j \neq i$ 有 Z_i 不变.

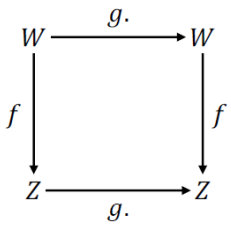

图 2 条件机制示意图.
The illustration of condition.

图 3 摆动的钟摆、光照与阴影.
Swinging pendulum, light and shadow.

定义 2 主要被基于变分自编码器的群论视角解耦表征学习方法所采用 (见第 3.1.1 节).

讨论. 以上两种定义都基于一个假设: 生成因子天然彼此独立. 然而, Suter 等人 [14] 提出从结构因果模型 (Structural Causal Model, SCM) [42] 的视角来重新定义解耦表征学习, 在该视角下, 他们在模型中引入一组额外的混淆变量, 这些混淆变量在因果上影响可观测数据的生成因子. Yang 等人 [11] 和 Shen 等人 [43] 进一步放弃了独立性假设, 认为生成因子之间可能存在潜在的因果结构. 例如, 如图 3 所示, 光源的位置和摆锤的角度共同决定了影子的长度与位置. 因此, 他们不再采用独立性假设, 而是使用结构因果模型作为先验, 用于刻画生成因子之间的因果关系. 我们将这类假设生成因子存在因果关系的研究称为因果解耦方法, 其将在第 3.4 节中详细讨论.

3 解耦表征学习的分类

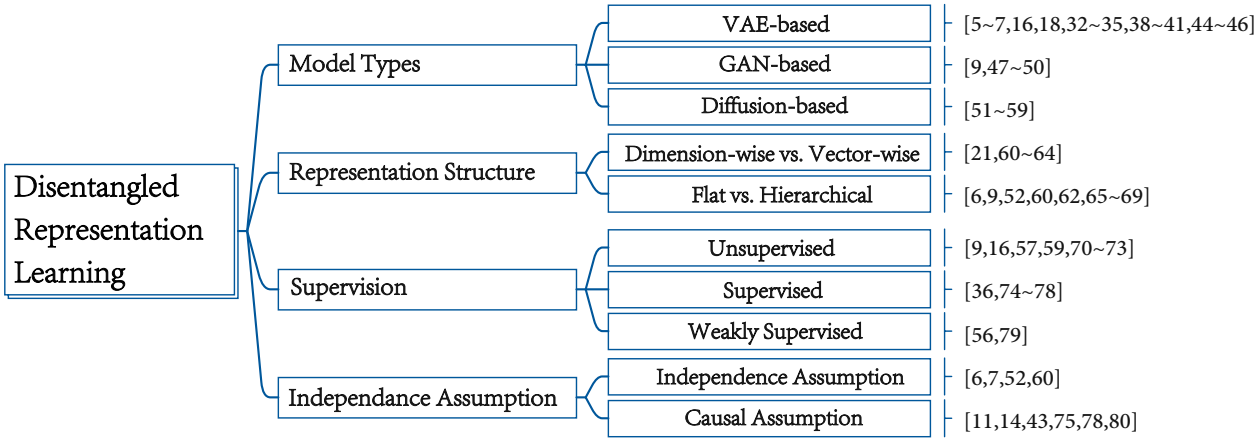

图 4 解耦表征学习方法的分类.
Categorization of DRL approaches.

如图 4 [5~7, 9, 11, 14, 16, 18, 21, 32~36, 38~41, 43~80] 所示, 我们从以下四个角度对解耦表征学习方法进行分类: (i) 基础模型类型; (ii) 表征结构类型, 即按维度划分与向量划分、平面结构与层级结构; (iii) 监督信号强度, 即无监督、弱监督与完全监督; (iv) 生成因子假设, 即独立性假设与因果性假设. 对于每一类方法, 我们将分别阐述其代表性模型、优势与局限, 并分析其典型的应用场景. 此外, 我们还将讨论胶囊网络与基于对象的学习, 因为这两种学习范式同样体现了解耦思想, 因此可视为解耦表征学习的特殊实例. 本文综述现有的解耦表征学习方法, 并探讨其在具体任务中的潜在应用与启示.

3.1 模型类型

在本节中, 我们根据底层模型架构对解耦表征学习方法进行分类. 在解耦表征学习研究的早期阶段, 研究者通常采用变分自编码器 (VAE) 或生成对抗网络 (GAN) 作为基础模型, 因此这类方法被称为基于 VAE 的或基于 GAN 的解耦表征学习方法, 它们可被视为传统方法. 近期的研究使解耦表征学习具备了更高的灵活性, 能够适应更先进的模型架构, 如扩散模型 (Diffusion Models). 我们希望通过展示解耦表征学习在传统模型和先进模型中的有效性, 来强调其优越性与泛化能力. 此外, 我们还将对这些模型进行一定的比较分析.

3.1.1 基于 VAE 的解耦表征学习

标准 VAE 基础方法. 变分自编码器是自动编码器的一种变体, 最早由 [16] 提出, 其核心思想是引入变分推断. VAE 最初被设计为一种深度生成概率模型, 用于生成复杂数据. 后续研究发现, VAE 具有在简单数据集

(如 FreyFaces [16] 与 MNIST [81]) 上学习到可分离潜在表征的潜力.

为了进一步提升解耦性能, 研究者设计了多种正则化策略, 将其与原始的 VAE 损失函数结合, 形成了一系列基于 VAE 的解耦表征学习方法. 基于 VAE 的方法中最具代表性的特征之一是潜在空间结构的 “维度解耦性”, 即潜在向量的不同维度对应不同的变化因素. 在第 3.2.1 节中, 我们将详细讨论维度解耦结构与向量解耦结构之间的比较.

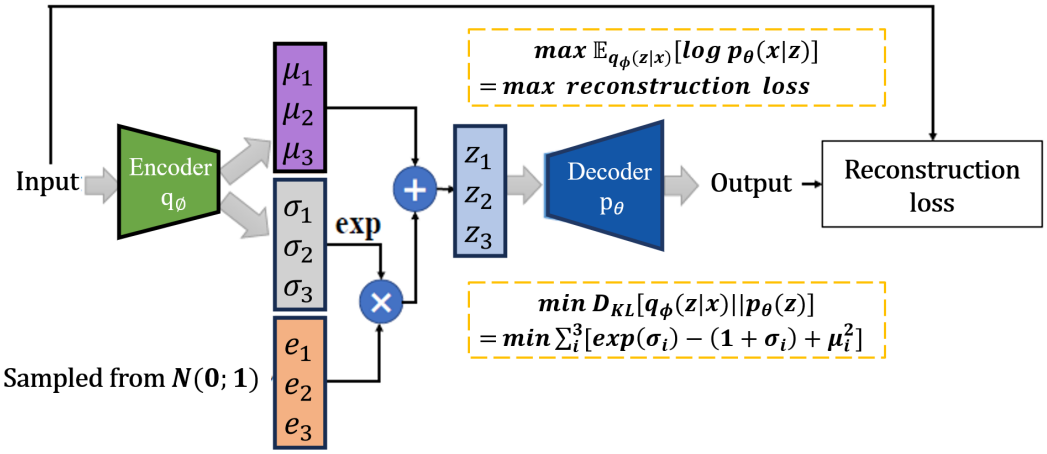

图 5 变分自编码器的一般框架.
The general framework of variational auto-encoder.

VAE 的一般模型结构如图 5 所示. 其基本目标是通过变分推断最大化似然估计, 即最大化 $\log p_\theta(x)$. 该目标可表示为式 (1):

$$\log p_\theta(x) = D_{KL}(q_\phi(z|x)||p_\theta(z|x)) + \mathcal{L}(\theta, \phi; x, z), \tag{1}$$

其中, q 表示近似的后验分布, z 为潜在变量. 式 (1) 的核心在于用可计算的变分后验分布 $q_\phi(z|x)$ 近似真实后验分布 $p_\theta(z|x)$, 后者在实践中往往难以直接求得. 更为详细的推导可参考原始文献 [16]. 式 (1) 的第一项是 $q_\phi(z|x)$ 与真实先验分布 $p_\theta(z)$ 之间的 KL 散度, 第二项被称为 (变分) 证据下界 (Evidence Lower Bound, ELBO). 在实际训练中, 我们通常最大化 ELBO, 以获得对原始似然 $\log p_\theta(x)$ 的紧下界. ELBO 可进一步写为式 (2):

$$\mathcal{L}(\theta, \phi; x, z) = -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)], \tag{2}$$

其中, 第二项 $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ 对应重建项, 而第一项 KL 散度项约束潜在分布接近于先验分布 $p_\theta(z)$. 通常, $p_\theta(z)$ 被选为标准高斯分布 $N(0, I)$, 这样 KL 散度项实际上就对通过神经网络学习得到的表征施加了独立性约束 [5]. 这也可能是变分自编码器具备潜在解耦能力的原因.

尽管 VAE 具有潜在的解耦能力, 但已有研究表明, 标准 VAE 在 CelebA [82] 和 3D Chairs [83] 等复杂数据集上往往表现出较差的解耦效果. 为改善这一问题, 大量研究提出通过显式或隐式地引入归纳偏置来增强解耦性能, 并设计了多种正则化变体, 如 β-VAE [6]、DIP-VAE [35] 与 β-TCVAE [5] 等. 具体而言, 为强化变分后验分布 $q_\phi(z|x)$ 的独立性约束, β-VAE [6] 在 ELBO 的 KL 散度项前引入了系数 β, 从而得到更新后的目标函数:

$$\mathcal{L}(\theta, \phi; x, z, \beta) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)||p_\theta(z)), \tag{3}$$

当 $\beta = 1$ 时, β-VAE 退化为原始的 VAE 形式. 实验结果表明, 较大的 β 值有助于学习到更解耦的潜在表征, 但同时会削弱重建质量 [6]. 因此, 在重建精度与解耦程度之间选择合适的 β 值至关重要. 为进一步解释这种权衡

关系, Chen 等人 [5] 从变分下界分解的角度提出了理论分析. 他们指出, 较大的 β 通常会增强潜变量间的维度独立性, 但同时限制潜在表征 z 保留输入 x 信息的能力.

然而, 在实践中, 找到最优的 β 值以平衡重建与解耦仍较为困难. Burgess 等人 [34] 将 β-VAE 重新解释为一种基于信息瓶颈理论的优化问题, 其目标函数如下:

$$\max[I(Z;Y) - \beta I(X;Z)], \tag{4}$$

其中, X 表示待压缩的原始输入, Y 表示目标任务, Z 是 X 的压缩表征, $I(\cdot\,;\cdot)$ 表示互信息. 回顾 β-VAE 框架, 我们可以将式 (4) 中的第一项 $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ 视作 $I(Z;Y)$, 并将第二项 $D_{\mathrm{KL}}(q_\phi(z|x)\|p_\theta(z))$ 近似视作 $I(X;Z)$. 具体来说, $q_\phi(z|x)$ 可以被看作重构任务 $\max \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ 的信息瓶颈; 而 $D_{\mathrm{KL}}(q_\phi(z|x)\|p_\theta(z))$ 可以看作 $q_\phi(z|x)$ 能从原始数据 x 中提取并保留的信息量的上界. 该策略的核心思想是逐步增加潜在通道的信息容量, 其修改后的目标函数如式 (6) 所示.

$$\mathcal{L}(\theta,\phi;C,\gamma) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \gamma[D_{KL}(q_\phi(z|x)\|p_\theta(z)) - C], \tag{5}$$

其中, γ 与 C 为超参数. 在训练过程中, C 将从 0 逐步增大, 以保证潜在空间的表达能力, 并在实现高质量解耦的同时维持合理的重建性能.

此外, DIP-VAE [35] 通过引入额外的正则化项进一步提高了解耦能力, 其优化目标定义如下:

$$\max_{\theta,\phi} \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p_\theta(z)) - \lambda D(\Sigma_{q_\phi(z)}\|I), \tag{6}$$

其中 $D(\cdot\|\cdot)$ 表示分布距离度量函数. 作者指出, 理想情况下 $q_\phi(z)$ 应等于 $\prod_i q_\phi(z_i)$ 以保证解耦性. 在假设 $p_\theta(z)$ 服从标准高斯分布 $\mathcal{N}(0,I)$ 的前提下, 目标函数对变分后验累积分布 $q_\phi(z)$ 施加了独立性约束. 为了最小化距离项, Kumar 等人通过在 $p_\theta(z) \sim \mathcal{N}(0,I)$ 的条件下, 使 $z \sim q_\phi(z)$ 的各维度去相关 (decorrelation), 以此匹配 $q_\phi(z)$ 与 $p_\theta(z)$ 的协方差. 换句话说, 他们强制式 7 尽可能接近单位矩阵.

$$\mathrm{Cov}_{q_\phi(z)}[z] = \mathbb{E}_{p(x)}[\Sigma_\phi(x)] + \mathrm{Cov}_{p(x)}[\mu_\phi(x)], \tag{7}$$

其中, $\mu_\phi(x)$ 与 $\Sigma_\phi(x)$ 分别表示后验分布 $q_\phi(z|x)$ 的均值与协方差预测. DIP-VAE 提出了两个变体: DIP-VAE-I 与 DIP-VAE-II, 其目标函数分别为:

$$\max_{\theta,\phi} \mathrm{ELBO}(\theta,\phi) - \lambda_{od}\sum_{i\neq j}[\mathrm{Cov}_{p(x)}(\mu_\phi(x))_{ij}]^2 - \lambda_d\sum_i([\mathrm{Cov}_{p(x)}(\mu_\phi(x))_{ii}]-1)^2, \tag{8}$$

$$\max_{\theta,\phi} \mathrm{ELBO}(\theta,\phi) - \lambda_{od}\sum_{i\neq j}[\mathrm{Cov}_{q_\phi(z)}[z]_{ij}]^2 - \lambda_d\sum_i([\mathrm{Cov}_{q_\phi(z)}[z]_{ii}]-1)^2, \tag{9}$$

其中, λ_d 与 λ_{od} 分别控制对角线与非对角线约束强度. 前者直接正则化均值协方差 $\mathrm{Cov}_{p(x)}[\mu_\phi(x)]$, 后者则直接约束潜变量协方差 $\mathrm{Cov}_{q_\phi(z)}[z]$.

Kim 等人 [7] 提出了 FactorVAE 方法, 其通过在目标函数中引入独立性正则项进一步约束潜变量的统计独立性 (详见式 (11)).

$$\frac{1}{N}\sum_{i=1}^{N}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})\|p_\theta(\mathbf{z}))\right] - \gamma D_{\mathrm{KL}}(q_\phi(\mathbf{z})\|\overline{q}_\phi(\mathbf{z})), \tag{11}$$

其中 $\overline{q}_\phi(\mathbf{z}) = \prod_j q_\phi(\mathbf{z}_j)$ 且 $\mathbf{x}^{(i)}$ 表示第 i 个样本. $D_{\mathrm{KL}}(q_\phi(\mathbf{z})\|\prod_j q_\phi(\mathbf{z}_j))$ 被称为总相关性, 用于评估 $\mathbf{z}$ 中各维度之间的独立程度.

Chen 等人 [5] 提出将 $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$ 详细分解为三项, 如式 (12) 所示. 其中, 第一项表示互信息, 可重写为 $I_q(\mathbf{z};\mathbf{x})$; 第二项表示总相关性; 第三项为逐维度的 KL 散度.

$$D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) = \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z},\mathbf{x})\|q_\phi(\mathbf{z})p_\theta(\mathbf{x}))}_{\text{(i) 互信息}} + \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z})\|\prod_j q_\phi(\mathbf{z}_j))}_{\text{(ii) 总相关性}} + \sum_j \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}_j)\|p_\theta(\mathbf{z}_j))}_{\text{(iii) 逐维度 KL 散度}}, \tag{12}$$

由式 (12) 可直接得到 β-VAE 中权衡关系的解释, 即: 增大 β 倾向于降低 $I_q(\mathbf{z};\mathbf{x})$, 这与重构质量相关; 同时增强 $q_\phi(\mathbf{z})$ 的独立性, 这与解耦相关. 因此, 与其对 $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$ 整体施加系数 β 进行惩罚, 不如分别对这三项施加不同的系数, 称为 β-TCVAE, 如式 (13) 所示.

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \alpha I_q(\mathbf{z};\mathbf{x}) - \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z})\|\prod_j q_\phi(\mathbf{z}_j)) - \gamma\sum_j D_{\mathrm{KL}}(q_\phi(\mathbf{z}_j)\|p_\theta(\mathbf{z}_j)). \tag{13}$$

为了进一步区分有意义的与噪声型的变化因素, Kim 等人 [33] 提出了 Relevance Factor VAE(RF-VAE) 模型, 引入了相关性指标变量, 用以衡量潜变量识别有意义变化因素的能力以及其数量. 上述基于 VAE 的方法多用于学习连续潜变量, 也有研究探索离散变量的情形. 例如, Dupont 等人 [32] 提出 β-VAE 框架的扩展版本 JointVAE, 该模型能够在无监督情境下同时学习连续与离散潜变量. 相关的数学公式在表 1 中补充.

表 1 基于 VAE 的方法总结.
The summary of VAE based approaches.

Methods	Regularization Terms	Description
β-VAE	$-\beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z}))$	β controls the trade-off between reconstruction fidelity and the quality of disentanglement in latent representations.
Understanding disentangling in β-VAE	$-\gamma\left\|D_{\mathrm{KL}}(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})) - C\right\|$	The quality of disentanglement can be improved as much as possible without losing too much information from original data by linearly increasing C during training.
DIP-VAE	$-\lambda D_{\mathrm{KL}}(q_\phi(\mathbf{z})\|p(\mathbf{z}))$	Enhance disentanglement by minimizing the distance between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$. In practice, we can match the moments between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$.
FactorVAE	$-\gamma D_{\mathrm{KL}}(q_\phi(\mathbf{z})\|\prod_j q_\phi(z_j))$	Directly impose independence constraint on $q_\phi(\mathbf{z})$ in the form of total correlation.
β-TCVAE	$-\alpha I_q(\mathbf{z};\mathbf{x})$ $-\beta D_{\mathrm{KL}}(q(\mathbf{z})\|\prod_j q(z_j))$ $-\gamma\sum_j D_{\mathrm{KL}}(q(z_j)\|p(z_j))$	Decompose $D_{\mathrm{KL}}(q(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z}))$ into three terms: i) mutual information, ii) total correlation, iii) dimension-wise KL divergence and then penalize them respectively.
JointVAE	$-\gamma\left\|D_{\mathrm{KL}}(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})) - C_z\right\|$ $-\gamma\left\|D_{\mathrm{KL}}(q_\phi(\mathbf{c}\|\mathbf{x})\|p(\mathbf{c})) - C_c\right\|$	Separate latent variables into continuous $\mathbf{z}$ and discrete $\mathbf{c}$, then modify the objective function of β-VAE to capture discrete generative factors
RF-VAE	$-\sum_{j=1}^{d}\lambda(r_j)D_{\mathrm{KL}}(q(z_j\|\mathbf{x})\|p(z_j))$ $-\gamma D_{\mathrm{KL}}(q(\mathbf{r}\circ\mathbf{z})\|\prod_{j=1}^{d}q(r_j\circ z_j))$ $-\eta\|\mathbf{r}\|_1$	Introduce relevance indicator variables $\mathbf{r}$ by only focusing on relevant part when computing the total correlation, penalize $D_{\mathrm{KL}}(q(\mathbf{z}_j\|\mathbf{x})\|p(\mathbf{z}_j))$ less for relevant dimensions and more for nuisance (noisy) dimensions.

综上所述, 上述基于 VAE 的无监督解耦表征学习方法的共同特征是引入额外的正则项 (例如 $D_{KL}(q_\phi(z)\|p_\theta(z))$ [35] 和总相关项 [7]), 以保证模型的解耦能力. 所有这些方法在 ELBO 基础上进行扩展, 其核心思想在表 1 中进行了总结. 此外, 还有若干研究将监督信号融入 VAE 框架中, 用以改进表征学习性能, 这部分将在第 3.3 节中进一步讨论.

VAE 还可以拓展用于处理时序数据, 如视频、音频等. Li 等人 [44] 将视频帧的潜在表征分为时间不变与时间变化两部分: 用 z_{t1} 表示视频帧的全局时间不变表征, 用 z_{t2} 表示时间变化表征. 其训练目标仍遵循 VAE 算法 (式 14):

$$\mathcal{L}(\theta,\phi,\mathbf{x},\mathbf{z},\mathbf{f}) = \mathbb{E}_{q_\phi(\mathbf{z}_{1:T},\mathbf{f}|\mathbf{x}_{1:T})}\left[\log p_\theta(\mathbf{x}_{1:T}\mid\mathbf{z}_{1:T},\mathbf{f})\right] - D_{KL}\Big(q_\phi(\mathbf{z}_{1:T},\mathbf{f}\mid\mathbf{x}_{1:T})\parallel p_\theta(\mathbf{z}_{1:T},\mathbf{f})\Big). \tag{14}$$

基于群论的 VAE 方法. 除了直观定义 (定义 1) 外, Higgins 等人 [18] 提出了一个数学上更为严格的基于群论的解耦表征学习定义 (定义 2), 并引发了一系列关于群结构表征学习的研究 [38~41].

Quessard 等人 [39] 提出了一种从动态环境 (返回观测序列) 的变换轨迹中学习解耦表征的方法. 他们考虑数据空间 O 和潜在表征空间 V, 其中数据集由轨迹 $(o_0, g_0, o_1, g_1, \dots)$ 构成, o_i 表示数据的观测, $g_i \in G$ 表示将 o_i 变换为 o_{i+1} 的变换. 他们将 G 映射为特殊正交群 $SO(n)$ 中的矩阵群, 即将 g_i 映射为一般线性群 $GL(V)$ 中的元素, 如式 (16) 所示. $R_{i,j}$ 表示在 (i,j) 平面上的旋转变换. 例如, 在三维情况下:

$$R_{i,j}(\theta_{i,j}) = \begin{pmatrix} \cos\theta_{i,j} & 0 & \sin\theta_{i,j} \\ 0 & 1 & 0 \\ -\sin\theta_{i,j} & 0 & \cos\theta_{i,j} \end{pmatrix}. \tag{15}$$

在训练过程中, 首先随机选择一条轨迹中的观测 o_i, 然后通过式 (17) 生成一系列重构 $\{\hat{o}_k\}_{k=i+1,\dots,i+m}$, 其中 f_ϕ 是将观测映射到 n 维潜在空间 V 的编码器, d_ψ 是解码器. 第一个目标是最小化重构损失 $\mathcal{L}_{\text{rec}}(\phi,\psi,\theta)$, 即真实观测 $\{o_k\}_{k=i+1,\dots,i+m}$ 与由潜在空间中的变换生成的观测 $\{\hat{o}_k\}_{k=i+1,\dots,i+m}$ 之间的差异. 此外, 为强制实现解耦, 他们提出另一个损失函数 $\mathcal{L}_{\text{ent}}(\theta)$, 用于惩罚变换 $g_a(\theta_{i,j})$ 中涉及的旋转数量, 如式 (18) 所示. 降低 $\mathcal{L}_{\text{ent}}$ 表明 g_a 涉及更少的旋转并作用于更少的维度, 从而实现更好的解耦.

$$g(\theta_{1,2},\theta_{1,3},\dots,\theta_{1,n},\theta_{2,3},\dots,\theta_{2,n},\dots,\theta_{n-1,n}) = \prod_{i=1}^{n-1}\prod_{j=i+1}^{n} R_{i,j}(\theta_{i,j}), \tag{16}$$

$$\hat{o}_{i+m}(\phi,\psi,\theta) = d_\psi\left(g_{i+m}(\theta)\cdot g_{i+m-1}(\theta)\cdots g_{i+1}(\theta)\cdot f_\phi(o_i)\right), \tag{17}$$

$$\mathcal{L}_{\text{ent}}(\theta) = \sum_a \sum_{(i,j)\neq(\alpha,\beta)} \|\theta^a_{i,j}\|^2 \quad \text{with} \quad \theta^a_{\alpha,\beta} = \max_{i,j}(|\theta^a_{i,j}|). \tag{18}$$

与依赖环境提供世界状态的环境驱动方法 [38,39] 不同, Yang 等人 [40] 提出了一个理论框架, 使定义 2 在无监督解耦表征学习的设定下可行, 而无需依赖环境. 该框架提出了三个充分条件: **模型约束**、**数据约束**和**群结构约束**, 并通过一个具体的基于现有 VAE 模型的实现, 将额外的辨识损失引入模型中. 作者假设群 G 是整数环模 n 的直积形式, 即

$$G = (\mathbb{Z}/n\mathbb{Z})^m = \mathbb{Z}/n\mathbb{Z}\times\cdots\times\mathbb{Z}/n\mathbb{Z},$$

其中 n 表示每个因子的可能取值数, m 表示因子总数. 假设 Z 与 G 拥有相同的元素, 并进一步假设 G 在 Z 上的群作用为加法形式, 即

$$g\cdot z = \overline{g+z}, \quad \forall z\in Z,\ g\in G,$$

为了避免在无监督场景中直接引入在世界状态空间 W 上的群作用, 他们构造了置换群 Φ, 并将 Φ 在数据空间 O 上的群作用替换为在 W 上的群作用. 该替换可由式 (19) 形式化定义如下:

$$f(g \cdot w) = h(\varphi_g \cdot b(w)) = h(\varphi_g \cdot o), \quad \forall w \in W, g \in G, \tag{19}$$

其中, h 表示从 O 到 Z 的映射, b 表示从 W 到 O 的映射.

满足式 (19) 的群作用 Φ 存在当且仅当以下条件成立:

1. Φ 与 G 同构;

2. 对于群 G 的每一个生成元 g_i, 存在对应的 Φ 的生成元 φ_i, 使得 $\varphi_i \cdot b(w) = b(g_i \cdot w), \forall w \in W$;

3. $\varphi_g \cdot b(w) = h^{-1}(g \cdot h(o)), \forall w \in W, \forall \varphi_g \in \Phi$, 其中 φ_g 是 g 在同构下的对应群元.

条件 1 与 2 分别被称为**群结构约束**和**数据约束**. 条件 3 是一个**模型约束**, 它进一步确保群 Φ 能够通过编码器、解码器与群作用在 Z 上得以实现. 当这三个条件均满足时, 可以推导出 Z 关于 G 是可解耦的. 然而, 鉴于条件 2 直接涉及世界状态), 研究者提出了一种名为 GROUPIFIED VAE 的学习方法, 该方法利用一个必要条件来替代条件 2, 以满足无监督学的要求. 因此, 在 VAE 的框架下, 条件 3 中的模型约束可以表示为:

$$\varphi_g \cdot o = h^{-1}(g \cdot h(o)) \triangleq d(g \cdot h(o)), \quad \forall o \in O, g \in G, \tag{20}$$

其中 h 为编码器, d 为解码器. 此外, 数据约束在某种程度上可由 VAE 模型在无监督设定下满足, 因为 VAE 模型能够生成与统计独立潜变量相似的样本. 为了进一步满足群结构约束, GROUPIFIED VAE 提出了 Abel Loss 与 Order Loss, 以确保 Φ 与 G 同构, 其定义如下:

$$\mathcal{L}_a = \sum_{o \in O} \sum_{(i,j)} |\varphi_i \cdot (\varphi_j \cdot o) - \varphi_j \cdot (\varphi_i \cdot o)|, \tag{21}$$

$$\mathcal{L}_o = \sum_{o \in O} \sum_{1 \leqslant i \leqslant m} (\|\varphi_i^{n-1} \cdot o - o\| + \|\varphi_i^{n-1} \cdot (\varphi_i \cdot o) - o\|), \tag{22}$$

其中 φ 为 Φ 的生成元.

在将同态映射从群扩展至群作用之后, Wang 等人 [41] 提出了迭代划分不变风险最小化 (Iterative Partition-based Invariant Risk Minimization, IP-IRM) 算法, 这是一种在自监督学习范式下构建的算法. 具体而言, 该方法旨在学习从观测空间 $\mathcal{I}$ 到表征空间 $\mathcal{X}$ 的映射, 即学习一个可解耦的表征提取器 f, 使得 $x = \phi(f)$, 并在群论基础的解耦学习条件下保持结构一致性. 他们首先指出, 大多数现有的自监督学习方法仅能对增强表征进行部分解耦, 而无法实现全局语义的模块化. 在这种背景下, IP-IRM 方法通过引入分组矩阵 $\mathbb{P}$ 将训练数据划分为若干不相交的子集, 并对第 k 个子集样本施加对比损失 $\mathcal{L}(\phi, \theta = 1, k, \mathbb{P})$, 其中 θ 为常数参数. 在每次迭代中, 方法通过最大化群内方差找到新的划分 $\mathbb{P}^*$, 由下式 (23) 给出, 以显式揭示一个可解耦的群元素 g_i:

$$\mathbb{P}^* = \arg\max_{\mathbb{P}} \sum_k \left[\mathcal{L}(\phi, \theta = 1, k, \mathbb{P}) + \lambda_2 \|\nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1, k, \mathbb{P})\|^2 \right]. \tag{23}$$

随后, 采用不变风险最小化 (Invariant Risk Minimization,IRM) 方法 [45] 进行更新, 如式 (24) 所示, 以在表征空间上实现关于群 g 的解耦. 具体而言, 初始设置 $\mathbb{P} = \{\mathbb{P}_i\}$, 并在每次迭代后更新 $\mathbb{P} \leftarrow \mathbb{P} \cup \mathbb{P}^*$:

$$\min_{\phi} \sum_{\mathbb{P} \in \mathcal{P}} \sum_k \left[\mathcal{L}(\phi, \theta = 1, k, \mathbb{P}) + \lambda_1 \|\nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1, k, \mathbb{P})\|^2 \right]. \tag{24}$$

理论上, 反复执行上述两个步骤将收敛至完全可解耦的表征, 即关于所有群生成元 $\prod_{i=1}^{m} g_i$ 的解耦表征. IP-IRM 的设计旨在推迟群作用学习至下游任务中按需触发, 从而通过一个推理过程学习到解耦表征, 在大规模任务上兼具可行性与可扩展性.

此外, Zhu 等人 [46] 提出了一种无监督强化学习框架, 称为 Commutative Lie Group VAE. 他们引入了一个矩阵李群 G 及其对应的李代数 $\mathfrak{g}$, 并定义其满足以下条件 (式 25):

$$\begin{aligned} &g(t) = \exp(B(t)), \quad g \in G,\ B \in \mathfrak{g}, \\ &B(t) = t_1 B_1 + t_2 B_2 + \cdots + t_m B_m, \quad \forall t_i \in \mathbb{R}. \end{aligned} \tag{25}$$

其中 $\exp(\cdot)$ 表示矩阵指数映射, $\{B_i\}_{i=1}^{m}$ 为李代数的基. 在此情形下, 每个样本可表示为一个群表征, 并可通过李代数的坐标来辨识. 该模型的优化目标由式 26 给出:

$$\begin{aligned} \log p(\mathbf{x}) &\geqslant \mathcal{L}_{\text{bottleneck}}(\mathbf{x}, \mathbf{z}, \mathbf{t}) \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})q(\mathbf{t}|\mathbf{z})} \log p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{t}) \\ &\quad - \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} D_{\text{KL}}(q(\mathbf{t}|\mathbf{z})\|p(\mathbf{t})) - \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log q(\mathbf{z}|\mathbf{x}), \end{aligned} \tag{26}$$

其中 $q(z|x)$ 实现为确定性编码器, $q(t|z)$ 实现为随机编码器, $p(z \mid t)$ 通过 $z = g(t)$ 实现, 而 $p(x \mid z)$ 则由图像解码器实现. 前两项分别表示基于数据空间和表征空间的重建损失. 第三项为条件熵, 其为常数项. 此外, 还提出了一类基于组合约束和 Hessian 惩罚约束的变体方法, 用以促进表征的解耦.

3.1.2 基于 GAN 的解耦表征学习

生成对抗网络 (GAN) [17] 是由 Goodfellow 等人提出的重要生成模型方法, 已受到广泛关注. 与传统的贝叶斯统计方法不同, GAN 直接从先验分布 $p(z)$ 中采样潜变量表征 z. 具体而言, GAN 包含生成器 G 与判别器 D, 其中生成器 G 通过潜变量 z 生成样本, 而判别器 D 则接收输入 (无论为真实样本或生成样本), 并输出其为真实的概率. 在训练过程中, G 的目标是生成以假乱真的样本, 从而 “欺骗” 判别器 D; 而 D 的目标则是区分真实与生成的样本. 因此, G 与 D 构成一个动态的极小极大博弈. 理想情况下, 当 G 能够生成与真实数据无法区分的样本时, D 无法判定其真假, 博弈达到平衡. 其目标函数定义如下 (式 27):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))], \tag{27}$$

其中 p_{data} 表示真实数据分布, $p(z)$ 表示潜表征的先验分布.

类似于基于 VAE 的方法, 研究者也提出了多种基于 GAN 的方法来实现解耦表征学习.

InfoGAN [9] 是最早利用 GAN 范式进行维度级解耦表征学习的方法之一. 生成器接受两个潜变量作为输入, 其中一个是不可压缩噪声 z, 另一个是目标潜变量 c, 用于捕捉潜在生成因素. 为了鼓励 c 的解耦, InfoGAN 设计了一个额外的互信息变分正则项, 即 $I(c; G(z,c))$, 由超参数 λ 控制, 使得 InfoGAN 的对抗损失可写为公式 28 :

$$\min_G \max_D V_I(D,G) = V'(D,G) - \lambda I(c; G(z,c)), \tag{28}$$

其中 $V'(D,G)$ 定义为式 29:

$$V'(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z,c)))]. \tag{29}$$

然而, 直接优化 $I(c;G(z,c))$ 在实践中难以实现, 因为后验分布 $p(c|x)$ 不可直接获取. 因此, InfoGAN 通过变分推断对互信息 $I(c;G(z,c))$ 给出下界, 如式 30 所示:

$$\begin{aligned} I(\mathbf{c};G(\mathbf{z},\mathbf{c})) &= H(\mathbf{c}) - H(\mathbf{c}|G(\mathbf{z},\mathbf{c})) \\ &= \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\left[\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})}[\log p(\mathbf{c}'|\mathbf{x})]\right] + H(\mathbf{c}) \\ &= \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}[\underbrace{D_{KL}(p(\cdot|\mathbf{x})||q(\cdot|\mathbf{x}))}_{\geqslant 0} + \mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})}[\log q(\mathbf{c}'|\mathbf{x})]] + H(\mathbf{c}) \\ &\geqslant \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\left[\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})}[\log q(\mathbf{c}'|\mathbf{x})]\right] + H(\mathbf{c}) \\ &= \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}),\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}[\log q(\mathbf{c}|\mathbf{x})] + H(\mathbf{c}). \end{aligned} \tag{30}$$

其中, $H(\cdot)$ 表示随机变量的熵, $q(c \mid x)$ 是近似真实后验分布 $p(c \mid x)$ 的辅助分布. 实际上, q 由神经网络实现. InfoGAN 的整体结构如图 6 所示.

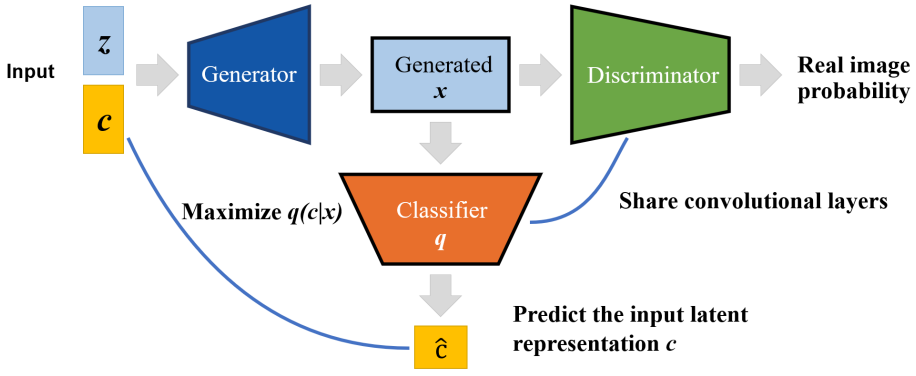

图 6 InfoGAN 的整体框架.
The overall framework of InfoGAN.

尽管 InfoGAN 在可解释性表征学习方面具有较好的理论基础, 但其在解耦表现上通常被认为弱于基于变分自编码器的模型. 为此, Jeon 等人 [47] 提出了 IB-GAN, 通过在潜变量表征 z 与生成结果 $G(z,c)$ 之间引入互信息最大化约束, 从而压缩表征空间. 该约束实质上起到了信息瓶颈的作用, 使潜表征更趋向于解耦.

Lin 等人 [48] 提出了 InfoGAN-CR, 这是一种基于对比正则化的自监督变体. 其方法在保持部分潜变量维度不变的情况下生成多张样本, 即随机固定某个维度 c_j, 采样其他维度 c_i $(i \neq j)$, 并训练分类器判断哪一维被固定. 此类对比正则化鼓励模型学习到不同维度的语义独立性, 从而促进解耦表征.

Zhu 等人 [49] 基于 InfoGAN 提出 PS-SC GAN, 该方法采用空间约束设计, 以获取每个潜在维度的关注区域, 并利用感知简洁性设计, 鼓励潜在表征捕捉的变异因子更简单、更纯粹. 空间约束设计通过空间掩码实现受限修改. 此外, PS-SC GAN 对特定潜在维度 c_i(即 $c_i' = c_i + \epsilon$) 施加扰动 ϵ, 然后计算 $\mathbf{c}$ 与 $\hat{\mathbf{c}}$ 之间的重构损失, 其中 $\hat{\mathbf{c}} = q(G(\mathbf{c},\mathbf{z}))$; 同时计算 $\mathbf{c}'$ 与 $\hat{\mathbf{c}}'$ 之间的重构损失, 其中 $\hat{\mathbf{c}}' = q(G(\mathbf{c}',\mathbf{z}))$, 这里 q 为 InfoGAN 中的分类器. 感知简洁性原则是对扰动维度的重构误差施加更大惩罚, 同时对剩余维度的错位给予更大容忍.

Wei 等人 [50] 提出正交雅可比正则化, 用于强制生成模型实现解耦. 该方法利用输出相对于输入 (即潜在表征变量) 的雅可比矩阵, 衡量输入变异引起的输出变化. 假设不同潜在维度引起的输出变化相互独立, 则雅可

比向量应相互正交, 即最小化式 (31):

$$\mathcal{L}_{\text{Jacob}}(G) = \sum_{d=1}^{D} \sum_{i=1}^{m} \sum_{j \neq i} \left| \left[\frac{\partial G_d}{\partial z_i} \right]^{\top} \frac{\partial G_d}{\partial z_j} \right|^2, \tag{31}$$

其中 G_d 表示生成模型的第 d 层, z_i 表示潜在表征的第 i 个维度.

3.1.3 基于扩散的解耦表征学习方法

一方面, 扩散模型是一类先进的生成模型, 近年来备受关注. 扩散模型在多种任务中表现出色, 包括文本到图像生成 [84]、图像编辑 [85] 以及文本到视频生成 [86] 等. 另一方面, 解耦表征学习也可应用于此类先进模型, 体现了解耦表征学习的广泛适用性. 因此, 本节将简要介绍扩散模型的定义, 并阐述若干基于扩散的解耦表征学习方法.

鉴于扩散模型通常涉及时间步, 本节采用下标表示不同时间步, 上标表示不同因子. 形式上, 扩散模型定义了一个前向过程, 向真实数据 $\mathbf{x}_0$ 逐步添加噪声:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{32}$$

其中 α_t 和 σ_t 为 t 的函数, 且信噪比 α_t/σ_t 严格递减. 扩散模型训练一个神经网络 $\boldsymbol{\epsilon}_\theta$ 来预测含噪变量 $\mathbf{x}_t$ 中的噪声, 其训练目标为:

$$\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, t} \left[\| \boldsymbol{\epsilon}_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, t) - \boldsymbol{\epsilon} \|_2^2 \right]. \tag{33}$$

在逆向过程中, 扩散模型通过逐步去噪从高斯噪声 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 生成样本, 直至恢复真实数据 $\mathbf{x}_0$, 即:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}),$$

其中 $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ 由扩散网络 $\boldsymbol{\epsilon}_\theta$ 结合特定采样轨迹确定 [87~89].

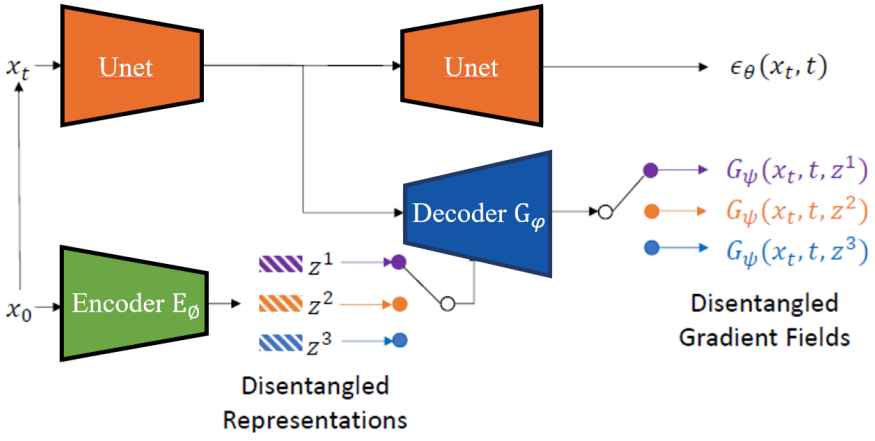

图 7 DisDiff 的网络结构.
The network structure of DisDiff.

扩散模型也可受益于解耦表征学习的优势. Yang 等人 [51] 提出无监督解耦框架 DisDiff, 其网络结构如图 7 所示. 他们学习每个生成因子的解耦表征以及解耦梯度场. 具体而言, 他们为每个因子 f^i 分配独立的编码器 E_ϕ^i, 以提取解耦表征 z^i, 即:

$$\{z^1, z^2, \ldots, z^N\} = \{E_\phi^1(x_0), E_\phi^2(x_0), \ldots, E_\phi^N(x_0)\}.$$

受类引导扩散模型采样 [90] 启发, 他们采用解码器 $G_\psi(\mathbf{x}_t, z, t)$ 估计得分函数, 即梯度场 $\nabla_{\mathbf{x}_t} \log p(z^i|\mathbf{x}_t)$. 采样过程可表述为:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) + \sigma_t \sum_i \nabla_{\mathbf{x}_t} \log p(z^i|\mathbf{x}_t), \sigma_t). \tag{34}$$

确保这些表征 z^i 之间解耦的关键在于, 作者提出了解耦损失, 通过最小化上界来抑制 z^i 与 $z^j (i \neq j)$ 之间的互信息:

$$\mathcal{L}_{\text{dis}} = \min_{i,j,x_0,\hat{x}_0^i} \|\hat{z}^{i|j} - \hat{z}^i\| - \|\hat{z}^{i|j} - z^i\| + \|\hat{z}^{j|j} - z^j\|, \tag{35}$$

其中 $\hat{x}_0^j$ 通过采样过程 $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, z^j)$ 条件于 z^j 生成, $\hat{z}^i = E_\phi^i(\hat{x}_0^j)$, $\hat{z}^{i|j} = E_\phi^i(\hat{x}_0^j)$. 注意, $\hat{x}_0$ 表示采样数据, 而 x_0 表示真实数据.

Chen 等人 [52] 提出了解耦的 stable diffusion [84] 微调框架 DisenBooth, 通过将表征解耦为身份保持部分和身份无关部分, 实现了主体驱动的文本到图像生成. 具体而言, 他们采用 CLIP 文本编码器提取文本身份保持文本嵌入, 并采用适配器提取视觉身份无关嵌入 (例如背景). 为确保解耦, 他们还设计了若干特定损失函数, 包括弱去噪损失和对比损失, 以进一步增强解耦性. 后续工作中 [53] 引入了样本特定掩码和样本特定去噪损失.

Chen 等人 [91] 进一步提出了解耦微调框架 VideoDreamer, 用于多主体驱动的视频生成. 他们将输入主体在单一图像中混合, 并利用三类解耦表征表示真实的三种独立成分, 即:i) 图像特定信息, 如背景和姿势;ii) 由混合引起的缝合条件;iii) 混合的主体身份信息.

3.1.4 预训练生成模型中的解耦潜在方向

大多数基于 GAN 和 VAE 的解耦表征学习方法都遵循从头开始搜索某些潜在方向的范式, 这些方向与某些语义上可解释的变换相关. 然而, 最近的工作 [92~94] 已经显示, 语义上更有意义的变化往往沿着预训练生成模型潜在空间中的某些方向进行. 这些现象表明, 存在某些属性在潜在空间中已经被预训练生成器解耦. 每个生成的潜在方向都与一个生成因子相关联. 发现解耦的潜在方向是一种更有效的方式来实现解耦表征学习, 相比于传统的从头开始的方法, 它可以利用预训练模型的强大能力并节省资源. 预训练模型在大数据时代扮演着越来越重要的角色, 尤其是在为大型模型配备解耦能力时, 通过发现解耦的潜在方向可能会成为一种有前景的路径. 在本节中, 我们将深入介绍几种代表性的工作, 从 GAN 到扩散模型.

Voynov 等人 [94] 提出了第一个用于发现预训练 GAN 潜在空间中解耦潜在方向的无监督框架. 具体来说, 他们在潜在空间 $\mathbb{R}^{d\times K}$ 中学习一个矩阵 A, 其中 d 表示潜在空间的维度, K 是总生成因子的数量. 矩阵 A 的第 k 列表示第 k 个因子的方向向量. 该方向向量可以被添加到潜在码 z 以变换对应的生成因素. 为了学习 A, 他们获取图像对的形式为 $(G(z), G(z + A(e_k))$, 其中 $z \sim \mathcal{N}(0, I)$, G 表示预训练生成器, e_k 表示单位向量, ε 表示一个标量偏移. 构造函数 R 被用来预测索引 k 并将给定的生成图像对 $(G(z), G(z + A(e_k))$ 生成到 $(G(z), G(z + A(e_k))$. 损失函数如下所示:

$$\min_{A,R,z,k,\varepsilon} L(A, R) = \min_{A,R,z,k,\varepsilon} \mathbb{E}\left[L_{cl}(k, \hat{k}) + \lambda L_r(\varepsilon, \varepsilon)\right], \tag{36}$$

其中 $\hat{k}, \varepsilon = R(G(z), G(z + A(e_k)))$. 最小化该损失将迫使 A 形成解耦的方向向量, 这些向量更容易被重建器 R 区分.

Ren 等人 [95] 提出了一种无监督框架 Disco, 用于对比学习以发现潜在方向. 类似于 Voynov 等人 [94], 他们也利用可学习的矩阵 $A \in \mathbb{R}^{d\times K}$ 来表示潜在方向, 其中每一列代表一个候选方向. 此外, 他们采用编码器 E 来显式提取解耦表征. 值得注意的是, 这里的 “解耦表征” 不是在预训练生成器 G 的原始潜在空间中, 而是在编

码器提取的独立空间中. 因此, 所发现的解耦潜在方向可用于解耦的可控性生成, 同时提取的解耦表征可应用于下游任务. 作者采用对比学习来联合优化潜在方向 A 和编码器 E. 具体来说, 他们首先构造一个变分空间, 由 $v(z,k,\varepsilon) = |E(G(z + A(e_k))) - E(G(z))|$, 其中 k 表示方向索引, ε 表示偏移量. 然后对比损失被用来将同一 k 的样本拉近并推开来自不同 k 的样本.

Kwon 等人 [54] 提出了一种名为 Asyrp 的框架, 通过在预训练扩散模型的最深层表征图空间中发现语义方向, 实现图像编辑. 具体而言, 他们将方向 Δh_i 与第 i 个属性 (如 “微笑”) 对齐, 其中 h 为预训练 UNet 的最深层表征图. 随后, 通过沿该方向平移实现属性编辑, 即 $\tilde{h} = h + \alpha\Delta h_i$. 该平移将改变 UNet 输出, 即预测噪声 ϵ_θ, 进而改变生成图像的第 i 个属性. 他们采用基于 CLIP 的方向损失 (使用余弦距离) 来优化 Δh_i.

此外,3.1.3 节中介绍的 DisDiff [51] 也可视为一种方向发现方法, 因为它在梯度场中获取了解耦的潜在方向 $\nabla_{\mathbf{x}_t} \log p(z^i|\mathbf{x}_t)$.

3.1.5 比较

正如我们在前文所述, 基于 VAE 的方法通常在可逆性、生成能力 (即解释能力) 和生成质量 (即重构性能) 方面存在权衡. 虽然基于 GAN 的模型通常具有显著的生成能力, 但它们缺乏可逆性属性, 这使得它们在诸如 VAE 等更灵活的模型中不太适用. 扩散模型能够结合 GAN 和 VAE 的优势. 扩散模型不仅具有强大的真实生成能力, 而且自然具备可逆性, 展现出成为未来解耦表征学习研究主流的巨大潜力.

表 2 基于 FactorVAE 分数和 DCI 指标的解耦性能比较 (均值 ± 标准差; 数值越高表示性能越好; 最优结果以粗体标出. 次优结果以下划线标出).
Comparison of disentanglement performance on FactorVAE score and DCI metrics (mean ± standard deviation; higher values indicate better performance; the best results are in bold, and the second-best results are underlined).

	Method	Shapes3D [4]		Cars3D [96]		MPI3D [97]	
		FactorVAE score↑	DCI↑	FactorVAE score↑	DCI↑	FactorVAE score↑	DCI↑
VAE based	FactorVAE [7]	0.840 ± 0.066	0.611 ± 0.082	0.906 ± 0.052	0.161 ± 0.019	0.152 ± 0.025	0.240 ± 0.051
	β-TCVAE [98]	0.873 ± 0.074	0.613 ± 0.114	0.855 ± 0.082	0.140 ± 0.019	0.179 ± 0.017	0.237 ± 0.056
	DAVA [99]	0.820 ± 0.030	0.780 ± 0.030	0.940 ± 0.010	0.230 ± 0.040	0.480 ± 0.050	0.270 ± 0.030
	DisCo-VAE [95]	0.956 ± 0.041	0.844 ± 0.033	0.761 ± 0.114	0.211 ± 0.041	0.391 ± 0.075	0.288 ± 0.021
GAN based	InfoGAN-CR [100]	0.587 ± 0.058	0.478 ± 0.055	0.411 ± 0.013	0.020 ± 0.011	0.439 ± 0.061	0.241 ± 0.075
	ClosedForm [92]	0.951 ± 0.021	0.525 ± 0.078	0.873 ± 0.036	0.243 ± 0.048	0.523 ± 0.056	0.318 ± 0.014
	GANSpace [101]	0.788 ± 0.091	0.284 ± 0.034	0.932 ± 0.018	0.209 ± 0.031	0.465 ± 0.036	0.229 ± 0.042
	DisCo-GAN [95]	0.877 ± 0.031	0.708 ± 0.048	0.855 ± 0.074	0.271 ± 0.037	0.371 ± 0.030	0.292 ± 0.024
Diffusion based	DisDiff-VQ [51]	0.902 ± 0.043	0.723 ± 0.013	**0.976** ± 0.018	0.232 ± 0.019	0.617 ± 0.070	0.337 ± 0.057
	FDAE [78]	0.987 ± 0.023	0.917 ± 0.038	0.918 ± 0.027	0.232 ± 0.418	<u>0.903</u> ± 0.030	<u>0.644</u> ± 0.031
	EncDiff [102]	<u>0.999</u> ± 0.000	**0.969** ± 0.030	0.773 ± 0.060	<u>0.279</u> ± 0.022	0.872 ± 0.049	**0.685** ± 0.044
	DyGA [103]	**1.000** ± 0.000	<u>0.938</u> ± 0.001	<u>0.941</u> ± 0.002	**0.414** ± 0.013	**0.930** ± 0.004	0.627 ± 0.002

为了对上述定性分析进行定量验证, 并全面评估模型在不同数据分布下的解耦与泛化能力, 我们选取了 Shapes3D [4]、Cars3D [96] 和 MPI3D [97] 三个标准合成数据集, 以及 CelebA [82] 真实世界数据集构建了统一的实验基准. 在评估指标方面, 针对合成数据集, 我们主要汇报了 FactorVAE 分数 (FactorVAE score) [7] 与 DCI [104] 指标; 而对于真实数据集, 我们则采用 TAD [105] 与 FID [106] 指标来综合衡量模型的解耦性能与生成保真度. 表 2 和表 3 汇总了各主流范式代表性方法在该基准下的性能表现. 具体而言, 我们沿用了 DisCo 等前沿工作的实验设置, 为所有评估模型配备了结构相同的编码器以提取解耦表征, 并将表征维度统一设定为 32. 此外, 我们确保了所有实验的数据预处理流程与训练策略完全一致, 而各模型的其余超参数则均严格遵循原论文中的默认设置, 从而在消除干扰变量的前提下实现公平的性能对比.

标准合成基准数据集上的定量分析. 如表 2 所示, 在合成数据集中, 基于扩散模型的方法整体在 FactorVAE 分数和 DCI 等核心解耦指标上普遍优于基于 VAE 和 GAN 的传统基线模型. 特别是在包含复杂物理渲染机制的 MPI3D 数据集上, 扩散模型 (如 FDAE、EncDiff、DyGA 等) 相较于其他两类方法展现出了显著的性能提升, 验证了该范式在捕捉独立生成因子方面的有效性与稳定性.

表 3 在真实世界数据集 CelebA 上. 基于 TAD 和 FID 指标的解耦性能比较 (均值 ± 标准差).
Disentanglement performance comparison using TAD and FID metrics on the real-world dataset CelebA (mean ± standard deviation).

Model	TAD	FID
β-VAE [6]	0.088 ± 0.043	99.8 ± 2.4
InfoVAE [107]	0.000 ± 0.000	77.8 ± 1.6
Diffae [108]	0.155 ± 0.010	22.7 ± 2.1
InfoDiffusion [109]	0.299 ± 0.006	23.6 ± 1.3
DisDiff [51]	0.305 ± 0.010	18.2 ± 2.1
EncDiff [102]	0.638 ± 0.008	14.8 ± 2.3
DyGA [103]	0.954 ± 0.024	12.0 ± 1.2

真实场景数据集上的定量分析. 如表 3 所示, 在特征更复杂的真实数据集 CelebA 上, 传统 VAE 方法 (如 β-VAE 和 InfoVAE) 的 TAD 极低且 FID 偏高, 难以兼顾解耦提取与高保真重构. 相比之下, 扩散模型展现出显著优势: Diffae 等早期扩散方法将 FID 大幅降至 25 以内, 有效提升了生成质量; 而最新的 EncDiff 与 DyGA 则在维持极低 FID(分别为 14.8 与 12.0) 的基础上, 使 TAD 指标分别升至 0.638 与 0.954. 这一结果表明, 扩散模型不仅在合成数据上表现稳健, 也能较好地适应真实数据分布, 有效缓解了传统范式中表征解耦与生成保真度难以平衡的难题.

总结. 上述定量结果表明, 扩散模型凭借其强大的生成先验与可逆性特征, 在一定程度上缓解了传统范式中解耦性能与高保真生成之间的内在博弈. 无论是在受控的合成基准还是高维复杂的真实场景中, 扩散模型均展现出了良好的解耦表征学习潜力与研究前景.

3.2 表征结构

我们详细讨论解耦表征学习在表征结构方面的术语, 包括 i) 维度级与向量级, 以及 ii) 平坦与层次化.

3.2.1 维度级解耦表征学习与向量级解耦表征学习

基于解耦表征的结构, 我们可以将解耦表征学习方法分为两类, 即维度级和向量级方法. 设 k_1 和 k_2 分别表征生成因子的总数. 对于维度级方法, 它们采用一组解耦的维度 $z = \{z_i, i = 0, k_1 - 1\}$, 其中 z 是整体表征, 每个单一维度 z_i(即一维标量) 表示一个细粒度生成因子. 相比之下, 向量级方法采用 k_2 个解耦向量 $z_i, i = 0, \ldots, k_2 - 1$ 来表示 k_2 个粗粒度因子, 其中维度级方法中的每个向量 z_i 等于或大于 2. 两种方法的比较如图 8 和表 4 所示. 维度级方法总是在合成和简单数据集上进行实验, 而向量级方法常被应用在真实世界任务 (如身份交换、图像分类、主体驱动生成和视频理解等). 合成和简单数据集通常具有细粒度的潜在因子, 因此适用于维度级解耦. 相反, 对于真实世界数据集的应用, 我们通常关注粗粒度因子 (例如身份和姿势), 这更适合向量级解耦. 维度级方法主要是对解耦表征学习的早期的理论探索, 并通常依赖于特定模型架构, 例如 VAE 或 GAN. 向量级方法则更加灵活, 例如, 我们可以设计任务特定的编码器来提取解耦向量, 并设计适当的损失函数来确保解耦.

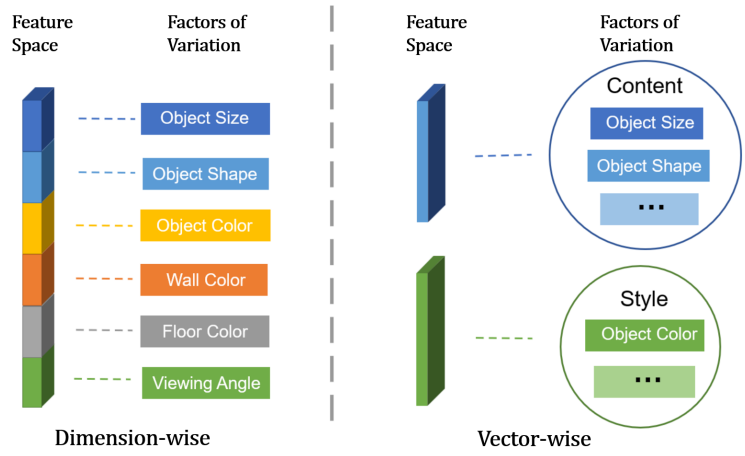

图 8 维度级与向量级解耦表征学习对比示意图.
The illustration of the comparison between dimension-wise and vector-wise DRL.

维度级和向量级方法之间的根本区别在于信息容量, 具体由维度的数量来表示. 具体来说, 维度级方法通常分配一个一维标量来表示一个生成因子, 而向量级方法以向量形式用更多维度表示生成因子. 随着维度数量的增加, 向量级表征自然能够捕捉更复杂的信息. 因此, 我们认为这是向量级解耦表征学习适用于粗粒度因子捕捉而维度级解耦表征学习适用于细粒度因子捕捉的主要原因. 然而, 我们注意到, 粒度有时是一个相对概念. 例如, 在简单的合成数据集上, "物体大小" 往往是一个细粒度因子, 足够用单一维度标量表征. 而另一方面, 在复杂的真实世界场景中, "物体大小" 可以被分解为 "宽度" 和 "高度" 等更细粒度的概念, 因此是一个粗粒度的概念, 需要更多维度来表示. 因此, 在实践中, 维度级或向量级方法的选择在很大程度上取决于任务的具体性以及目标场景和任务. 在本节中, 我们将讨论一系列维度级和向量级解耦表征学习方法的代表性工作.

表 4 维度级方法与向量级方法的比较.
The comparisons of dimension-wise and vector-wise methods.

Methods	Dimension of Each Latent Factor	Representative Works	Semantic Alignment	Applicability
Vector-wise	Multiple	MAP-IVR [55], DRNET [56], DR-GAN [57], DRANet [21], Lee et al. [8], Liu et al. [22], Singh et al. [58]	Each latent variable aligns to one coarse-grained semantic meaning	Real scenes
Dimension-wise	One	VAE-based methods, InfoGAN [9], IB-GAN [47], Zhu et al. [19], InfoGAN-CR [48], PS-SC GAN [49], Wei et al. [50], DNA-GAN [59]	Each dimension aligns to one fine-grained semantic meaning	Synthetic and simple datasets

维度级方法. 基于 VAE 的方法是一个可用于实现维度级解耦的典型架构, 我们在第 3.1.1 节中已花费大量篇幅讨论. 此外, 正如我们在第 3.1.2 节中讨论的那样, 基于 GAN 的方法是另一个重要的实现维度级解耦表征学习的范式. 接下来, 我们将更多地关注各种向量级方法.

向量级方法. 除了维度级解耦表征学习之外, 基于 GAN 的模型也可以用于在更真实的任务上实现向量级解耦表征学习. Tran 等人 [57] 提出了 DR-GAN, 用于姿态不变的人脸识别. 他们使用两个向量分别表示身份和姿态. 具体而言, 他们显式地设置一个独热 (one-hot) 潜变量向量来表示姿态, 并使用一个编码器从输入图像中提取身份向量, 同时分别针对姿态和身份设计了两个判别器, 以确保潜变量向量能够与相应的语义 (即生成因素) 对齐. 最终, 所学习到的身份向量可用于实现姿态不变的人脸识别, 或用于合成保持身份一致的人脸图像.

Liu 等人 [55] 提出了一种解耦框架 MAP-IVR, 用于活动图像到视频检索, 该框架把视频表征分离为外观和运动部分. 具体来说, 他们使用两个视频编码器从视频表征 $\tilde{v}$ 中分别提取视频运动表征 m^v 和视频外观表征 a^v . 他们还使用一个图像编码器来提取参考图像的整体外观表征. 他们设计了几个目标来确保解耦:

$$\mathcal{L}_{orth} = \cos(m^v, a^v), \tag{37}$$

$$\mathcal{L}_{class} = -\log(p(a^v)_y) - \log(p(a^v)_y), \tag{38}$$

$$\mathcal{L}_{re} = \|\bar{v} - \hat{v}\|_2^2. \tag{39}$$

$\mathcal{L}_{orth}$ 是一种余弦距离损失, 用于促进视频运动表征 m^v 和外观表征 a^v 之间的正交性. $\mathcal{L}_{class}$ 利用一个活动分类器 p 来确保视频外观表征 a^v 和图像外观表征 a^u 都能捕捉活动信息. 重构损失 $\mathcal{L}_{re}$ 也有助于解耦, 其中 $\bar{v}$ 是原始视频表征, $\hat{v}$ 是通过结合 a^u 与从 m^v 中得到的运动信息后重建得到的. 解耦的表征通过将图像参考转换为视频参考, 可用于实现更好的活动图像到视频检索.

Denton 等人 [56] 提出了一种基于自编码器的模型 DRNET [56], 其将每个视频帧解耦为时间不变 (内容) 和时间变化 (姿态) 组件. 他们使用两个编码器分别提取第 t 帧的内容表征 h_c^t 和姿态表征 h_p^t. 内容表征和姿态表征的解耦目标分别为:

$$\mathcal{L}_{re}(D) = \|D(h_c^t, h_p^{t+k}) - x^{t+k}\|_2^2, \tag{40}$$

$$\mathcal{L}_{sim}(E_c) = \|E_c(x^t) - E_c(x^{t+k})\|_2^2. \tag{41}$$

$\mathcal{L}_{re}(D)$ 是重构损失, 旨在确保组合 h_c^t 和 h_p^{t+k} 可以重构帧 x^{t+k}. $\mathcal{L}_{sim}$ 表示 $h_c^i = E_c(x^t)$ 应该跨时间 t 不变. 这两个损失期望 h_c 捕捉时间不变的内容以及 h_p 捕捉时间变化的姿态. 此外, 他们还使用对抗损失来帮助 h_p 不携带内容相关的信息. 解耦的表征可被用于未来帧预测或分类任务.

Cheng 等人 [60] 提出 DFR 框架, 利用解耦表征进行少样本图像分类. 他们使用两个编码器 E_{cls} 和 E_{var} 分别提取类特定和类无关表征. 解耦目标为:

$$\mathcal{L}_{dis} = -\sum_{i=1}^{P}(l_i \cdot \log(s_i) + (1 - l_i) \cdot \log(1 - s_i)), \tag{42}$$

$$\mathcal{L}_{cls} = -\sum_{i=1} y_i \log P(\hat{y} = y_i \mid \mathcal{T}_{FS}), \tag{43}$$

$\mathcal{L}_{dis}$ 是一种判别损失, 用于移除类无关表征中的类特定信息. $s_i = r_\varphi(E_{var}(x_{i1}), E_{var}(x_{i2}))$, 其中 r_φ 是判别器, 输出 x_{i1} 和 x_{i2} 来自同一类的概率. 他们采用梯度反转层来鼓励 E_{var} 移除类特定信息. $\mathcal{L}_{cls}$ 是用于分类损失的交叉熵

损失, 其中预测 $\hat{y}_i$ 是基于类特定表征 $E_{cls}(x_i)$ 获得的. $\mathcal{L}_{cls}$ 鼓励 E_{cls} 捕捉类特定信息. 此外, 它们还采用重构损失和变换损失以进一步促进解耦.

Lee 等人提出了一种解耦跨域适应框架 DRANet [21]. 他们使用一个编码器提取内容表征, 而通过从原始表征中减去内容表征来获得风格表征. 他们还设计了几个损失, 例如感知损失, 以增强解耦. 域适应可以通过将内容表征与目标域风格表征组合来实现. Gao 等人提出了一种解耦的身份交换框架 InfoSwap [61], 该框架将身份相关表征与身份无关表征进行解耦. 他们通过优化基于信息瓶颈理论的损失目标来实现这种解耦. 身份交换可通过将身份相关表征与目标的身份无关表征相结合来实现.

讨论. 我们已经介绍了一系列维度级和向量级解耦表征学习方法. 维度级解耦表征学习方法使用单一维度 (或多个维度) 来表示一个细粒度生成因子, 而向量级解耦表征学习方法使用单个向量来表示一个粗粒度生成因子. 这两类方法的共同关键点是如何通过设计特定的损失目标 (例如, 各种正则化项或专门设计的监督信号) 来强制实现解耦. 我们将在第 6 节中深入讨论如何为解耦表征学习任务设计损失目标. 至于如何确定使用维度级还是向量级方法, 这取决于及生成因子的数量和粒度, 我们将在本节中重点讨论. 例如, 在许多真实世界应用中, 我们只需要考虑两个或几个粗粒度因子. 另一方面, 对于某些在简单数据集上的特定生成任务, 我们需要考虑多个细粒度因子. 维度级方法主要是在简单场景下对解耦表征学习的早期的理论探索, 而近年来, 研究者们更加关注如何将向量级解耦表征学习应用于真实世界应用中. 探索向量级解耦表征学习在各种真实任务中的潜力, 可能成为未来的发展趋势.

3.2.2 扁平解耦表征学习与分层解耦表征学习

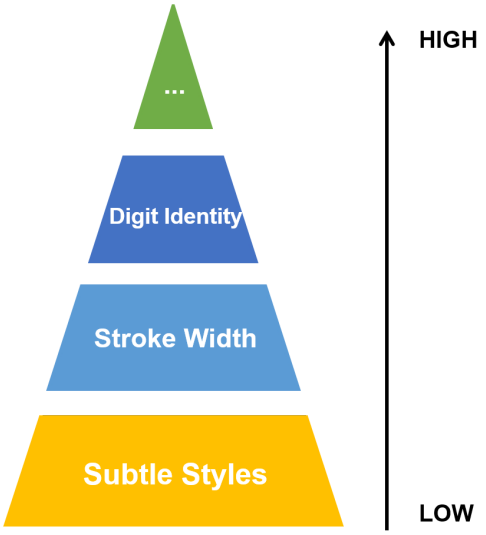

图 9 层次化解耦表征学习示意图. 生成因素之间存在层次结构. 即不同因素属于不同的抽象层级. 整体类似金字塔结构.
The illustration of hierarchical DRL. There exists a hierarchical structure among generative factors, i.e., the factors belong to different abstraction levels, resembling a pyramid.

上述解耦表征解耦方法普遍假设生成因子的结构是扁平的, 即所有因子彼此平行, 处于相同的语义抽象层次. 例如, 在维度级解耦表征学习方法中, β-VAE [6] 能够在 CelebA 数据集上解耦面部旋转、微笑、肤色、刘海等因子;InfoGAN [9] 则在 3D Faces 数据集上解耦方位角、仰角、光照等因子. 在向量级的解耦表征学习方法中, DR-GAN [57] 实现了人脸身份与姿态的解耦; MAP-IVR [55] 对视频中的运动表征与外观表征进行了解

耦;DisenBooth [52] 则解耦了与身份保持相关和与身份无关的表征. 总的来说, 这些被解耦的因子之间并不存在层级结构.

然而, 在实际的生成过程中, 生成因子往往天然具有层级结构 [62, 63], 其中因子在不同语义层次上具有不同的抽象程度, 这些层次之间的因子可能相互依赖 [62], 也可能相互独立 [63]. 例如, 在 CelebA 数据集中, 控制“性别” 的因子具有比控制 “眼影” 的因子更高的抽象层级, 且二者相互独立 [63]; 而在 Spaceshapes 数据集中, 控制 “形状” (高层因素) 与 “相位” (低层因素) 的因子存在依赖关系 [62]——例如, “相位” 维度仅当物体形状为 “月亮” 时才处于激活状态.

为了刻画这些层级化结构, 研究者提出一系列工作以实现分层解耦表征学习. 图 9 展示了分层解耦表征学习的基本范式.

Li 等人 [63] 提出了一种基于 VAE 的模型, 通过将分层生成概率模型公式化来学习分层解耦表征, 如式 (44) 所示:

$$p(x, z) = p(x \mid z_1, z_2, \ldots, z_L) \prod_{l=1}^{L} p(z_l), \tag{44}$$

其中 z_l 表示第 l 层的潜在表征, l 越大表示抽象级别越高. 作者估计抽象级别与网络深度相关, 即更深的网络层负责输出更高抽象级别的表征. 值得注意的是, 式 (44) 假设不同层次的潜在表征之间不存在依赖性. 换句话说, 每个潜在表征倾向于捕捉单一抽象层次中的因子, 这在其他层次中不会被覆盖. 相应的推理模型表述为式 (45):

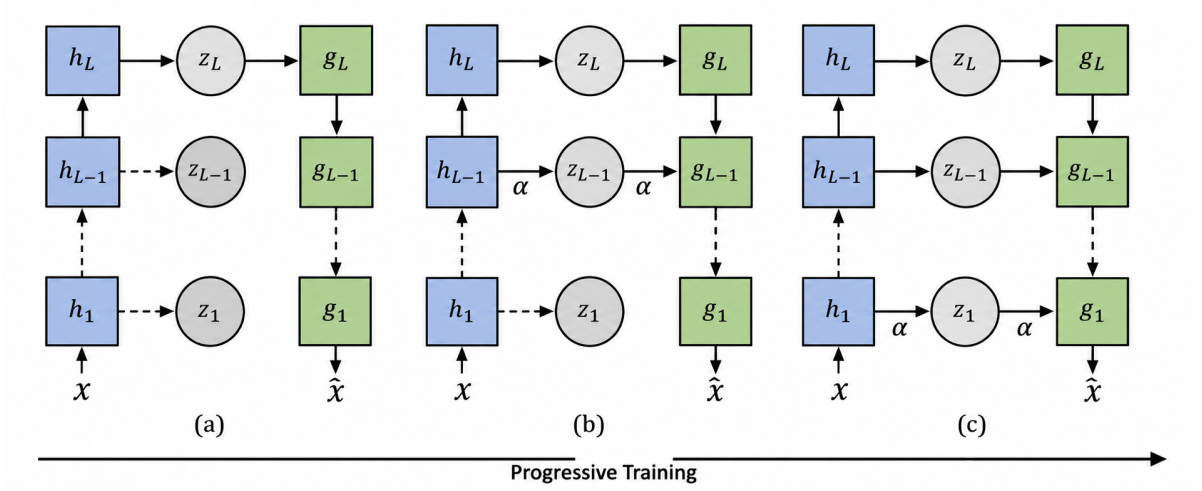

图 10 层次化框架架构.
The architecture of the hierarchical framework.

$$q(z_1, z_2, \ldots, z_L \mid x) = \prod_{l=1}^{L} q(z_l \mid h_l(x)), \tag{45}$$

其中 $h_l(x)$ 表示第 l 层的抽象. 在训练阶段, 作者设计了一种渐进式学习策略, 从高到低抽象级别逐步修改 ELBO 目标. 分层渐进式学习如图 10 所示, 其中 h_l 和 g_l 是一组编码器和解码器, 分别用于不同抽象级别. 该框架能够在 MNIST 数据集上, 从高到低的语义抽象层次中依次解耦数字身份、笔画粗细以及细微的数字风格表征; 同时, 它还能够在 CelebA 数据集上解耦性别、微笑、卷发以及眼影等属性.

Tong 等人 [64] 提出学习一组分层的解耦表征 $z = \{z_l^i\}_{i=1}^{c_l}$, 其中 z_l^i 是第 l 层的第 i 个潜在变量, c_l 是该层潜在变量的总数量. 为了确保在每个层次的解耦, 他们设计了如式 (46) 所示的损失函数:

$$\mathcal{L}_{\text{disentangle}} = \sum_{l} \frac{2}{c_l(c_l-1)} \sum_{i\neq j}^{c_l} d\text{Cov}^2(z_l^i, z_l^j), \tag{46}$$

其中 $d\text{Cov}^2(\cdot,\cdot)$ 表示距离协方差.

Singh 等人 [58] 提出了一种用于细粒度对象生成的无监督分层解耦框架——FineGAN. 他们为不同的层次级别设计了三个潜变量表征: 即背景编码 b、父级编码 p 和子级编码 c, 分别对应背景、对象形状与对象外观. 其中, 背景属于最低层次, 其次是形状, 最后是外观. 在生成过程中, FineGAN 首先以 b 和噪声 z 为输入, 生成一个逼真的背景图像; 随后, 以 p 和噪声 z 为输入生成对象的形状, 并将其叠加到背景图像之上; 最后, 以在父级编码 p 条件下的子级编码 c 作为输入, 模型将对象轮廓 (父级) 填充为包含外观细节 (子级) 的完整图像. 此外, 作者进一步借助信息论方法 (类似于 InfoGAN) 来实现父级 (形状) 与子级 (外观) 的解耦, 并结合对抗损失与辅助背景分类损失共同约束背景生成的质量.

Li 等人 [65] 提出了一种用于图像到图像翻译的分层解耦框架. 他们将标签按照语义抽象层次从高到低、从根节点到叶节点的顺序人工组织成层次树结构. 例如: 标签层 (如眼镜)、属性层 (如是否佩戴)、风格层 (如近视眼镜、太阳镜). 值得注意的是, 该树状层次结构表明子节点依赖于其父节点. 作者训练了一个翻译器模块以处理标签信息, 同时训练一个编码器以提取风格表征.

Ross 等人 [62] 同样提出了一种分层解耦框架, 其假设某些维度组仅在特定情况下才会被激活. 具体而言, 他们将生成因子组织为一个层次结构 (例如树状结构), 其中子节点是否激活取决于父节点的取值. 以 Spaceshapes 数据集为例, 当父节点 “形状” 的取值为 “moon” 时, 表示 “相位” 的维度才会被激活. 他们设计了一种名为 MIMOSA 的算法, 用于训练自编码器以学习分层解耦表征.

讨论. 综上所述, 我们可以根据生成因子是否具有层次结构来选择扁平或分层解耦表征学习方法. 虽然扁平解耦表征学习方法在理论上也有潜力去解耦不同语义抽象层次的因素, 但专门针对层次结构设计的分层解耦表征学习方法通常表现更优. 在具体应用中, 我们可以考虑是否存在可利用的潜在层次结构, 以此来促进解耦表征的学习.

3.3 监督信号

在本节中, 我们从学习机制的角度重新审视解耦表征学习方法, 即无监督与有监督的解耦表征学习. 这一问题既是广受关注的基础议题, 也是实现解耦表征的关键难点.

3.3.1 无监督方法

解耦表征学习方法的主要研究范式主要源于无监督学习, 早期研究尤其关注于自动发现可解释的、可分解的潜在表征. 原始的变分自编码器模型 [16] 展示了利用贝叶斯推断在无监督条件下学习潜在空间的可能性. 以 InfoGAN [9] 为代表的一类对抗生成网络模型, 致力于以无监督方式学习可解释的表征. 此外, 基于信息论的研究还提出了诸如互信息估计 [66] 和 DeepVIB [67] 等方法, 以解耦潜在信息并提升模型的鲁棒性与泛化能力.

作为解耦表征学习发展的初始阶段, 无监督学习范式被视为大多数研究者最初的研究愿景, 并代表了一类直观且有效的解耦表征学习实现途径.

3.3.2 监督方法

Locatello 等人 [68] 近期证明了: 若在方法与数据集上缺乏归纳偏置, 则 “纯粹的无监督解耦表征学习在理论上是不可能的” . 换言之, 解耦并不会自然发生, 这一结论打破了研究者长期以来专注于 “无监督解耦” 的局

面. 随后, Locatello 等人 [69] 提出, 在训练过程中引入部分有标签数据, 在解耦效果和下游任务性能两方面均具有积极作用.

DC-IGN [70] 约束在每个 mini-batch 中仅一个因素可变, 而其他因素保持不变. 在潜在表征 $\mathbf{z}$ 的各个维度中, 选择其中一维作为 z_{train}, 并通过监督训练使其能够解释该 batch 内的全部变化, 从而与被选定的可变因素对齐. MLVAE [36] 根据某一选定因素 f_s 将样本划分为若干组, 每组样本共享相同的 f_s 取值. 这种设定更适用于某些任务, 例如图像到图像翻译, 其中每组样本不仅具有相同的标签, 还共享关于 f_s 的潜在变量后验分布, 该后验依赖于组内所有样本. 而对于除 f_s 外的其他因素, 其后验分布则可能依赖于每个个体样本.

此外, Bengio 等人 [71] 指出, 从因果推理的角度来看, 模型的适应速度可用于评估模型是否契合潜在的因果结构, 并进一步提出了一种元学习目标函数, 以在未知因果变量混合的条件下学习具有良好表达性、可解耦且结构化的因果表征.

Xiao 等人 [59] 提出了 DNA-GAN, 一种训练过程类似于基因交换的监督模型. 具体而言, DNA-GAN 以具有不同标签的两张多标签图像 I_a 和 I_b 作为编码器输入. 通过编码器获得其原始表征 $\mathbf{a}$ 和 $\mathbf{b}$ 后, 在属性相关部分中交换某一特定维度的取值, 构造出交换后的表征 $\mathbf{a}'$ 与 $\mathbf{b}'$. 在解码阶段, 通过重建损失与对抗损失共同约束, 确保属性相关部分的每个维度都能够与相应的标签对齐.

此外, 来自重建损失与任务损失的监督信号, 也可视为一种广泛应用于实际场景的监督方式. 我们将在第 6 节 “解耦表征学习的设计” 中对此进行详细讨论.

3.3.3 弱监督方法

监督解耦表征学习方法要求目标数据集是语义清晰. 且结构良好的以便能够被解耦为可解释的、独立的、可恢复的生成因子 [110]. 然而, 在某些情形下, 可能存在一些不可处理或难以标注的潜在因素, 它们通常被视为与目标任务无关的噪声.

Xiang 等人 [110] 提出了一种弱监督的解耦表征学习框架 DisUnknown, 其设定为在总共 N 个因素中, 有 $N-1$ 个因素带有标签, 而剩余的 1 个因素未知. 因此, 所有不可处理的因素或与任务无关的因素都可以被归纳到这一单一的未知因素中. DisUnknown 模型采用两阶段方法, 包括: (i) 未知因素蒸馏; (ii) 多条件生成. 在第一阶段中, 模型通过对抗训练提取未知因素; 而在第二阶段, 模型将所有带标签的因素嵌入用于重建. 该方法利用一组判别分类器预测因素标签的概率分布以强化解耦, 这一思路与 InfoGAN [9] 类似.

Wang 等人 [76] 将弱监督解耦扩展到扩散模型框架. 由于扩散模型的生成过程是随机且不可逆的, 完全无监督的解耦同样无法实现, 因此作者引入部分标签或多视角信息作为弱监督信号, 并结合信息正则化的得分匹配目标, 使模型在学习数据分布的同时区分内容与风格等潜在因素. 该工作证明了在温和条件下, 扩散模型可在弱监督下恢复近似可辨识的解耦结构, 并提出相应的风格引导损失以进一步提升效果.

3.3.4 解耦表征学习的可识别性

解耦表征学习中最重要的关注点之一是可识别性 [11, 68, 72]. 这一问题主要在无监督解耦表征学习的背景下被讨论, 其核心关注点是无监督解耦表征学习的可行性. 可识别性指的是: 我们是否能够将期望获得的解耦模型与其他耦合的模型区分开来. Locatello 等人 [68] 指出, 如果在学习方法和数据集上都不引入归纳偏置, 那么通过无监督学习是无法识别出解耦模型的; 换句话说, 没有归纳偏置的无监督解耦表征学习是不可能的. 具体而言, 我们考虑如下生成范式:

$$p(x) = \int p(x|z)p(z)dz. \tag{47}$$

无监督解耦表征学习方法可以观测到 x, 即 $p(x)$, 但它无法根据边缘分布 $p(z)$ 识别真实先验潜在变量 z. 理论上, 存在无穷多个分布可以产生相同的边缘分布 $p(z)$ [68, 72]. 例如, 如果 $p(z)$ 是多变量高斯分布, 它对于旋转是不变的. 因此, 根据式 (47), 我们将得到无穷多个等价生成模型, 它们具有相同的 $p(z)$. 设 $\hat{z}$ 表示另一种生成模型的潜在变量, 则有:

$$p(x) = \int p(x|z)p(z)dz = \int p(x|\hat{z})p(\hat{z})d\hat{z}, \tag{48}$$

其中 $p(z)$ 等于 $p(\hat{z})$. 总之, 我们无法确保能获得解耦模型, 而不是其他等价的模型.

因此, 我们需要额外的归纳偏置来识别解耦模型, 或者利用监督信号来帮助找到目标模型 [68]. 对于前文提到的无监督方法而言, 它们之所以能够在一定程度上实现解耦表征学习, 是因为它们同样包含了归纳偏置, 例如正则项及其正则化强度. Locatello 等人 [68] 还指出, 无监督解耦表征学习的解耦评分容易受到随机性和超参数的影响. 因此, 尽管设计合适的归纳偏置对于无监督解耦表征学习非常重要, 但采用显式或隐式的监督信号可能更加有效.

3.4 独立性假设

直观而言, 前文讨论的典型解耦表征学习模型通常假设潜在因子之间是统计独立的, 因此可以通过独立或分解的正则化项 [6, 7] 或各种解耦损失函数 [52, 55] 来实现独立的解耦. 然而, 在某些情况下, 潜在生成因子并非相互独立, 而是存在一定的因果关系. 在本节中, 我们将讨论因果解耦表征学习方法, 该方法能够捕捉数据生成过程中的潜在因果机制, 并通过解耦因果因子来实现更具可解释性和鲁棒性的表征.

基于 Suter 等人 [14] 的论述, Reddy 等人 [111] 提出了生成潜变量模型 (例如 VAE) 应当满足的两个关键性质, 以实现因果解耦. 设一个潜变量模型为 $M(e, g, p_X)$, 其中 e 表示编码器, g 表示生成器, p_X 表示数据分布. 令 G_i 表示第 i 个生成因子, C 表示因果学习文献中所称的混淆变量 [112]. 关于编码器与生成器的这两个性质如下所示:

性质 1. 编码器 e 能够学习从生成因子 G_i 到唯一潜变量子集 Z_I 的映射, 其中 I 是一个索引集合, Z_I 是由 I 所索引的潜变量维度集合. "唯一" 意味着对于任意 $I \neq J$, 都有 $Z_I \cap Z_J = \emptyset$, 且 $|I|, |J| \geqslant 0$. 在这种情况下, 我们认为潜变量 Z 关于混淆变量 C 是无混淆的, 即对任意 $I \neq J$, Z_I 与 Z_J 之间不存在虚假相关.

性质 2. 对于生成器 g 所定义的生成过程, 只有 Z_I 能够影响由生成因子 G_i 所控制的生成输出部分, 而其他潜变量 (记为 Z_{I^-}) 则不能产生影响.

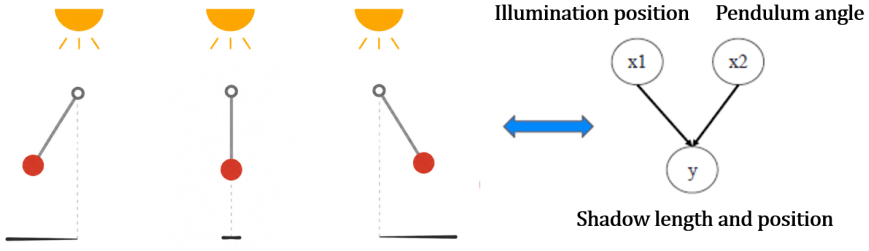

图 11 光源的位置和钟摆的角度是阴影位置和阴影长度的原因.
The position of the illumination source and the angle of the pendulum are causes of the position and the length of the shadow.

由于 Locatello 等人 [68] 质疑基于标准 VAE 的解耦表征学习方法中潜变量之间必须相互独立的常见假设, 随后有一些研究尝试放弃这一独立性假设. Yang 等人 [11] 提出了 CausalVAE, 该模型首次引入了结构因

果模型 (Structural Causal Model, SCM) 作为先验. CausalVAE 从因果关系的角度考虑数据中变化因素之间的相互关系, 并使用结构因果模型来描述这些关系, 如图 11 所示. CausalVAE 使用一个编码器将输入 x 以及与真实因果概念相关的监督信号 u 映射到一个独立的外生变量 ε, 其先验分布服从标准多元高斯分布 $\mathcal{N}(0, I)$. 这一编码过程如公式 (49) 所示.

$$\epsilon = h(x, u) + \zeta, \tag{49}$$

其中 h 是编码器, ζ 是噪声. 然后他们设计了一个因果层, 通过线性结构公式 (50) 将 ϵ 转化为因果表征 z:

$$z = A^\top z + \epsilon = (I - A^\top)^{-1}\epsilon, \tag{50}$$

其中 A 是因果有向无环图的可学习的邻接矩阵. 在送入解码器之前, z 通过一个掩码层进行重建, 如式 (51) 所示, 对于 z 的第 i 个隐空间维度 z_i:

$$z_i = g_i(A_i \circ z; \eta_i) + \epsilon_i, \tag{51}$$

其中 $\circ$ 表示逐元素乘积, g_i 是一个轻量级非线性函数, 参数为 η_i. 在掩码阶段, 通过将 z_i 的设置为固定值, 以 “执行操作 (do operation)” 的形式在进行因果干预. 掩码层之后, z 被送入解码器重建观测 x, 即 $\hat{x} = d(z) + \xi$, 其中 ξ 也是噪声.

Bengio 等人 [71] 指出, 从因果推断的角度来看, 适应速度可以用来评估一个模型与潜在因果结构的契合程度. 他们进一步利用元学目标函数, 在混合因果变量未知的情况下, 学习解耦且具备结构化的因果表征.

与 CausalVAE 的监督学习方案不同, Shen 等人 [43] 提出了一种名为 DEAR 的弱监督框架, 同样引入了结构因果模型作为先验. 首先, 通过编码器 E(或从先验分布 p_z 中采样) 获得因果表征 z, 即以样本 x 作为输入得到 $z = E(x)$. 其次, 根据 Yu 等人 [113] 提出的一般非线性结构因果模型, 计算外生变量 ε, 其中先前计算得到的 z 被用于定义 $F_\beta(\varepsilon)$, 如公式 (52) 所示.

$$[z = f_1((I - A^\top)^{-1} f_2(\epsilon))] := F_\beta(\epsilon), \tag{52}$$

其中 f_1 和 f_2 是逐元素变换, 它们常常是非线性的. A 是和式 (50) 和 (51) 一样的可学习邻接矩阵. β 表示 f_1、f_2 和 A 的参数. 当 f_1 可逆时, 式 (52) 等价于式 (53):

$$f_1^{-1}(z) = A^\top f_1^{-1}(z) + f_2(\epsilon). \tag{53}$$

第三步, 我们可以对 z “执行操作 (do operation)”, 即将 z_i 设为固定值, 然后通过迭代执行公式 (53) 进行祖先采样以重建 z. 最后, 将 z 传入解码器以实现重建. 为了保证解耦, 引入了一个弱监督损失函数 $L = \mathbb{E}_{x,y}[L_s(E; x, y)]$, 该方法仅需要少量标注数据. 其中, 当标签 y_i 被二值化时, $L_s = \sum_{i=1}^m \text{CrossEntropy}(\bar{E}(x_i), y_i)$; 当 y_i 为连续值时, $L_s = \sum_{i=1}^m (\bar{E}(x_i) - y_i)^2$. 注意, $\bar{E}$ 表示编码器 $E(x)$ 的确定性部分. 当采用 VAE 结构时, $\bar{E}(x) = m(x)$, 且 $E(x) \sim \mathcal{N}(m(x), \Sigma(x))$, 其中 $m(x)$ 和 $\Sigma(x)$ 分别表示编码器输出的均值与方差.

讨论. 在解耦表征学习中, 因果模型通常包含以下两个主要组成部分:

- 结构因果模型: 结构因果模型提供了一种利用有向无环图表示变量之间因果关系的方式. 在结构因果模型中, 图中的每个节点代表一个变量, 有向边表示因果依赖关系. 这些变量可以是可观测的, 也可以是潜变量, 模型明确描述了变量之间如何相互作用.

- 干预推理: 因果模型支持干预操作, 即我们可以对特定变量进行操作或干预, 从而观察其他变量的变化. 干预通常涉及改变模型中特定变量的取值, 并观察其对其他变量造成的影响. 这有助于理解变量之间的因果关系, 并将因果机制编码进解耦表征中.

通过将因果模型引入解耦表征学习, 我们可以显式地识别并解耦影响观测数据的因果因素. 在我们看来, 与独立假设下的解耦表征学习相比, 因果解耦表征学习方法更适用于存在多个生成因素及其潜在因果关系的场景. 需要注意的是, 学习因果模型与解耦表征是一个具有挑战性的任务. 在实际应用中, 准确地刻画潜在因果结构并捕获数据中的全部因果关系往往十分困难. 为应对这些挑战, 研究者通常采用因果推断、因果发现以及结构方程建模等多种技术与算法.

3.5 解耦表征学习的相互关系

在本节中, 我们讨论解耦表征学习与胶囊网络和基于对象的学习这两个学习范式之间的相互关系, 这两种范式在学习解耦表征方面具有密切关系, 可以被视为解耦表征学习的特定实例.

3.5.1 胶囊网络的相互关系

胶囊网络 (Capsule Networks) [114~117] 由 Hinton 等人 [115] 提出, 是传统卷积神经网络 (CNN) 的一个替代方案, 旨在解决 CNN 的一些局限性, 例如池化操作带来的信息损失, 以及无法处理局部-整体空间关系的问题. 与标量神经元单元相比, 胶囊网络将神经元组织为胶囊, 每个胶囊由一组神经元组成, 共同表示特定表征, 如姿态、颜色和纹理. 胶囊不仅能够编码对象的存在, 还能表示对象的属性以及与其他对象的关系. 胶囊网络通过多层胶囊显式地建模部分-整体的层次关系. 低层胶囊捕获低抽象层次的信息, 高层胶囊捕获高抽象层次的信息. 信息通过路由策略从低层胶囊传递到高层胶囊.

胶囊网络的概念本质上与解耦表征学习的思想相契合, 因为它倾向于将各种变化因素表示为独立的胶囊, 并将对象的表征分解为组成部分. 然而, 无法确保每个胶囊的表示确实是解耦的 [118]. 因此, 获得解耦胶囊仍然是一个开放性问题. Hu 等人 [118] 提出了 beta-CapsNet, 通过引入信息瓶颈约束来学习解耦胶囊. 实验结果表明, beta-CapsNet 在学习解耦表征的能力上优于 beta-VAE 和原始 CapsNet. 我们认为, 胶囊网络本身具有通过其层次化的部分-整体架构实现解耦表征学习的潜力, 为学习解耦表征提供了合适的框架. 此外, 还需要额外的正则化项或显式监督来确保胶囊的解耦. 胶囊网络与解耦表征学习的结合, 有望获得更具可解释性和鲁棒性的数据信息表征, 毫无疑问, 这是一个非常有前景的研究方向.

3.5.2 与基于对象的学习的相互关系

传统的深度学习方法通常将场景作为整体进行处理, 而不显式考虑单个对象及其组成结构. 相比之下, 基于对象的学习 (object-centric learning) [119~122] 旨在对不同对象进行显式建模和理解, 并捕捉场景的潜在结构以及对象之间的关系. 例如, 基于对象的学习可以输出不同对象的掩码以及可用于下游任务的对象中心表征. 基于对象的学习强调将对象视为独立实体进行识别与推理的重要性, 这对可控图像生成 [123]、图像分割 [124] 以及视觉问答 [125] 等多种下游任务都有益. 此外, 还有一系列研究实现了更细粒度的基于对象的学习, 即进一步解耦每个对象的表征. 例如, Ferraro 等人 [126] 对每个对象的形状和姿态进行解耦; Li [127] 对每个对象的多个因素进行解耦, 如旋转和颜色. 总之, 基于对象的学习可以被视为解耦表征学习的一个特例, 其重点在于解耦单个对象及其属性的表征.

3.6 关于分类之间联系的讨论

我们从四个方面展示了解耦表征学习的分类法: (i) 模型类型, (ii) 表征结构, (iii) 监督信号, 以及 (iv) 独立性假设, 并为每个方面列举了大量相关工作. 尽管这些研究可能关注解耦表征学习的不同方面, 我们认为这四个方面是相互关联的, 其共同假设是数据中存在可解释的潜在因素, 而最终目标是发现这些因素. 例如, 独立性假设会影响模型设计: 当潜在因素之间存在因果关系时, 我们可能选择因果 VAE 模型而非标准 VAE. 同样, 表征结构也会对模型类型产生影响, 例如维度级表征结构更适合基于 VAE 的模型. 未来的研究可以同时考虑这四个方面, 并在它们之间进行联合优化, 以在各种目标任务上实现有效的解耦表征学习.

4 度量

许多工作 [5~7,9] 通过观察在潜在空间中沿某一变量变化时重建结果的变化, 来定性评估解耦表征的性能. 定性观察直观易懂, 但不够精确, 也缺乏数学上的严格性. 为了推动解耦表征学习的研究, 设计能够定量衡量解耦性的可靠指标显得尤为重要. 我们回顾并将一系列定量指标分为两类: 监督指标和无监督指标. 有关指标的更深入理解、讨论与分类, 感兴趣的读者可参阅 Zaidi 等人的工作 [128].

4.1 监督度量

监督指标假设我们可以获得真实的生成因子.

Z-diff. Higgins 等人 [6] 提出了一种基于低容量线性分类器网络的监督解耦指标, 用于衡量独立性和可解释性. 他们在一组图像对上进行推断, 这些图像对是通过固定某个数据生成因子的值, 同时随机采样其他所有因子生成的. 以一批图像对作为输入, 分类器需要识别哪个因子被固定, 并将分类准确率作为解耦指标的得分.

Z-min Variance. Kim 等人 [7] 指出, 上述使用线性网络的方法存在若干缺陷, 例如对线性分类器优化的超参数敏感. 更重要的是, 该指标存在一种失败模式: 即使 K 个因子中只有 $K-1$ 个因子被解耦, 仍会给出 100% 的准确率. 作者提出了一种基于多数投票分类器的指标, 不需要优化超参数. 方法同样生成一组图像, 使其中一个因子 k 固定, 其余因子随机变化. 在获取归一化表征后, 选择经验方差最小的维度索引和标签 k 作为多数投票分类器的输入与输出. 分类器的准确率即被视为解耦指标得分.

Z-max Variance. Kim 等人 [33] 提出了一种指标, 与 Z-min Variance 基本相同. 主要区别在于, 他们生成样本时固定所有因子, 仅让一个因子 k 变化. 因此, 他们选择经验方差最大的维度索引作为多数投票分类器的输入. 他们认为该指标与定性评估的结果比 Z-min Variance 更一致.

MIG. Chen 等人 [5] 提出了一种无分类器的基于信息论的指标, 称为互信息间隙 (Mutual Information Gap, MIG). MIG 的核心思想是评估潜变量 z_j 与真实因子 k 之间的经验互信息. 对于每个因子 k, MIG 计算互信息最高的前两个潜变量之间的差距. 对所有因子取平均差距作为解耦指标得分. MIG 分数越高, 说明解耦性能越好, 因为这表明每个生成因子主要被单个潜在维度捕获.

SAP **分数**. Kumar 等人 [35] 提出了一种指标, 称为分离属性可预测性 (Separated Attribute Predictability, SAP) 得分. 他们构建了一个得分矩阵 $S \in \mathbb{R}^{d\times k}$, 其中第 (i,j) 元素表示仅使用第 i 个潜变量分布预测第 j 个生成因子的线性回归或分类得分. 然后, 对于得分矩阵的每一列, 他们计算前两大元素的差值, 并将这些差值的平均值作为 SAP 得分. SAP 得分越高, 表示解耦性能越好, 因为它也表明每个生成因子主要对应于唯一的潜变量维度, 类似于 MIG.

DCI. Eastwood 等人 [129] 设计了一个框架, 从三个方面评估解耦模型, 即解耦性 (Disentanglement, D)、完整性 (Completeness, C) 和信息性 (Informativeness, I). 具体来说, 解耦性表示每个潜变量最多捕获一个生

成因子的程度; 完整性表示每个生成因子仅被一个潜变量捕获的程度; 信息性表示潜变量捕获生成因子信息的量. 值得注意的是, 解耦性和完整性共同量化了双射映射与实际映射之间的偏差.

Modularity **和** Explicitness. Ridgeway 等人 [130] 从模块化 (Modularity) 和显性 (Explicitness) 两个方面评估解耦. 他们认为, 潜变量维度在理想情况下只有在与唯一一个因子具有高互信息且与其他因子互信息为零时才是模块化的. 模块化得分通过计算经验情况与理想情况的偏差获得. 显性侧重于潜变量表征对生成因子的覆盖程度. 假设生成因子取离散值, 他们使用一对多逻辑回归训练因子分类器, 并记录 ROC 曲线下面积 (AUC), 最终取所有因子所有类别的 AUC 平均值作为显性得分.

UNIBOUND. Tokui 等人 [131] 提出 UNIBOUND, 通过偏信息分解 (Partial Information Decomposition, PID) 中的唯一信息下界来评估解耦. PID 将潜变量 z_l 与生成因子 y_k 之间的信息分解为冗余信息、唯一信息和补充信息. 设 $\mathcal{U}(y_k; z_\ell \setminus \mathbf{z}_{\setminus \ell})$ 表示 z_l 拥有而其他变量不拥有的唯一信息, 他们对该唯一信息项进行下界估计. 类似于 MIG, 对于每个生成因子, 计算前两大潜变量的唯一信息下界差值, 并取所有生成因子的平均值作为最终得分.

UC **和** CG. Reddy [111] 从因果角度提出了无混淆性 (Unconfoundedness, UC) 指标和反事实生成性 (Counterfactual Generativeness, CG) 指标. 如第 2 节所述, 他们利用结构因果模型描述数据生成过程. UC 指标评估映射 $G_i \to Z_I$ 在给定混淆变量集合 C 下的唯一性和无混淆程度. UC 定义为

$$UC := 1 - \mathbb{E}_{x \sim p_X}\left[\frac{1}{|S|}\sum_{I,J}\frac{|Z_I^x \cap Z_J^x|}{|Z_I^x \cup Z_J^x|}\right].$$

CG 指标评估对 Z_I 的任何因果干预是否会影响生成 G_i 的相关方面, 这意味着生成过程中仅 Z_I 的干预会影响 G_i. CG 定义为

$$CG = \mathbb{E}_I\Big[|\mathrm{ACE}_{Z_I^X}^{X_I^{cf}} - \mathrm{ACE}_{Z_{\setminus I}^X}^{X_{\setminus I}^{cf}}|\Big],$$

其中

$$\mathrm{ACE}_{do(Z=\alpha)}^X = \mathbb{E}[X \mid do(Z=\alpha)] - \mathbb{E}[X \mid do(Z=\alpha^*)].$$

4.2 无监督度量

当无法获得真实的生成因子时，基于监督信息的解耦度量方法往往难以直接使用，因此无监督度量显得尤为重要。这类方法通常不依赖真实因子标签，而是从潜变量本身及其与观测数据之间的统计关系出发，评估表示的解耦程度。

ISI **指标**. Do 等人 [132] 从互信息的角度提出了一种用于评价解耦表示的指标体系，即 ISI 指标。该指标体系将解耦表示的性质概括为三个方面：信息性（Informativeness, I）、可分性（Separability, S）和可解释性（Interpretability, I）。其中，信息性衡量原始数据 x 与潜变量 z_i 之间所包含的互信息，可表示为 $I(x, z_i)$；可分性衡量任意两个潜变量 z_i 和 z_j 是否共享关于数据 x 的公共信息，即潜变量相对于数据 x 的可分能力，可表示为 $I(x, z_i, z_j)$；可解释性衡量潜变量 z_i 与数据生成因子 y_k 之间是否存在一一映射关系，可表示为 $I(z_i, y_k) = H(z_i) = H(y_k)$。此外，Do 等人还提出了相应的互信息估计方法，使该指标体系既可用于监督场景，也可用于无监督场景。

5 解耦表征学习的应用

在本节中, 我们讨论解耦表征学习在各种下游任务中的广泛应用. 首先, 我们将介绍解耦表征学习在不同模态 (例如图像、视频、自然语言以及多模态) 上的应用, 这些模态几乎涵盖了深度学习的所有应用领域. 此外,

我们还介绍解耦表征学习在推荐系统和图学习中的应用, 并在表 5 中给出简要总结. 最后, 我们展示解耦表征学习在小样本和分布外问题上的有效性.

表 5 解耦表征学习应用的代表性示例.
Representatives of disentangled representation learning applications.

Papers	Model	Paradigm	Application
[5~7, 16, 32~35]	VAE-based	Unsupervised	Image generation
[9, 10, 19, 133]	GAN-based		
[36, 69, 70, 134]	VAE-based	Supervised	
[57, 59]	GAN-based		
[11, 43]	Causal-based		
[8, 20~22, 135]	GAN-based	Unsupervised	Image translation
[136]	VAE-based	Supervised	Image classification,segmentation,etc.
[137, 138]	Others	Unsupervised	
[56, 139]	VAE-based	Unsupervised	Video
[55]	Others	Supervised	
[23]	Others	Unsupervised	Natural language processing
[13, 25]	Others	Supervised	
[140~142]	VAE-based	Unsupervised	Multimodal Application
[143, 144]	VAE-based	Supervised	
[145]	GAN-based		
[146~148]	Others		
[26, 28, 149~151]	VAE-based	Supervised	Recommendation
[27, 29, 152]	Others		
[153~156]	VAE-based	Supervised	Graph
[29, 30, 157]	Others		

5.1 图像

作为研究最广泛的视觉数据类型之一, 图像在生成、翻译、解释等任务中都能从解耦表征学习中获益良多.

5.1.1 生成

通过利用解耦表征学习, 可以在生成目标中学习和对齐独立因素与潜在表征之间的关系, 从而实现对生成过程的控制.

一方面, 最初的 VAE [16] 模型在图像生成与重建任务上能够学习良好的可解耦表征. 后续工作通过改进解耦与重建能力, 在图像操控与干预方面取得了更显著的效果. 代表性模型如 β-VAE [6, 34] 和 FactorVAE [7] 能更好地解耦变化的独立因素, 使得潜变量在图像生成过程中的可控操纵成为可能. JointVAE [32] 同时关注连续与离散表征, 相较于之前的方法获取了更具泛化性的表征, 从而将图像生成的应用范围扩展到更广的领域. CausalVAE [11] 引入弱监督下的因果结构, 使得生成具有因果语义的图像及构建反事实结果成为可能.

另一方面, 基于 GAN 的可解耦模型也被应用于图像生成任务, 并受益于 GAN 的高保真度. InfoGAN [9] 作为典型的基于 GAN 的模型, 以无监督方式解耦潜在表征以学习可解释的表征, 并在操控条件下生成图像, 但其稳定性与样本多样性较弱 [6, 7]. Larsen 等人 [10] 将 VAE 和 GAN 结合为一种无监督生成模型:i) 将解码器与生成器合并,ii) 使用表征级相似度而非逐元素误差,iii) 证明无监督训练能够产生一定的可解耦图像表征, 从而在高保真的图像生成与重建中学习高层视觉属性. Zhu 等人 [19] 利用 GAN 框架解耦三维表征 (如形状、视角、纹理), 从而合成自然物体图像. Wu 等人 [133] 分析了 StyleGAN [158] 中特别是 StyleSpace 的可解耦生成操作, 并操控语义上有意义的属性. Zeng 等人 [135] 提出混合模型 DAE-GAN, 结合可变形自动编码器与条件生成器, 从视频帧中解耦身份与姿态表征, 以无监督方式生成具有特定姿态的真实人脸图像.

其他基于信息论的工作也长期做出重要贡献. 例如, Gao 等人提出 InfoSwap [61], 通过优化信息瓶颈来解耦身份相关与身份无关信息, 以生成更加身份区分性的换脸结果.

5.1.2 域适应

除了生成任务之外, 域适应也是图像处理与理解中的热点方向. 解耦表征学习通过分离潜在的变化因素, 有助于在跨域场景中获得一致而稳健的性能, 从而增强并拓展域适应的可控性与适用性. 解耦表征学习通常用于解耦领域不变因素与领域特定因素, 以辅助域适应任务.

Gonzalez 等人 [20] 提出跨域解耦表征, 通过基于 GAN 的双向图像翻译以及跨域自编码器, 将内部表征解耦为共享部分与专属部分, 并仅以成对图像作为输入. 该设计在多样样本生成、跨域检索、域特定图像迁移与插值等任务上均取得了令人满意的表现. Lee 等人 [8] 在 GAN 框架中引入内容判别器与跨循环一致性损失, 将潜在表征解耦为领域不变的内容空间与领域特定的属性空间, 从而无需使用对齐图像对即可实现多模态多样性翻译. 随后, DRANet [21] 被提出用于解耦内容与风格因素, 并通过迁移视觉属性实现无监督的多方向域适应. Liu 等人 [22] 指出现有图像翻译模型在跨域与域内语义插值时缺乏渐变性. 因此他们提出新的训练协议, 以学习平滑且可解耦的潜在风格空间, 从而实现逐步变化并更好地保持源图像的内容.

5.1.3 其他应用

解耦表征学习的思想也广泛应用于其他图像相关领域与任务. Sanchez 等人 [137] 通过优化互信息来解耦成对图像中的共享与专属表征, 从而无需依赖图像重建或图像生成即可应用于图像分类和图像检索任务. Hamaguchi 等人 [136] 提出一种基于 VAE 的网络, 在仅需要成对观测的情况下, 对不平衡数据集中的罕见事件检测任务进行可变与不变因素的解耦. Gidaris 等人 [159] 通过让卷积网络预测图像旋转角度来进行自监督语义表征学习, 其性能可与监督方法相当. 受其启发, Feng 等人 [138] 在图像旋转预测与实例判别的联合训练中, 进一步将与语义旋转相关与无关的表征解耦, 从而提升图像分类、检索、分割等任务的泛化能力. Ghandeharioun 等人 [12] 提出 DISSECT 方法, 通过鼓励不同概念之间的差异性以及同一概念内部的紧致性来实现潜在概念的解耦. 该方法能够通过改变某些解耦概念来干预模型输出, 从而实现多种反事实图像解释.

总而言之, 通过各种图像表征学习模型获得的表征, 都可以基于解耦表征学习策略进行结构化, 从而将变化事件与固有属性进行分离. 因此, 作为图像相关任务的一种合适学习策略, 解耦表征学习特别有助于提升图像生成与图像翻译能力, 为多种图像应用带来更全面与多样的实现方式.

5.2 视频

除了静态图像之外, 解耦表征学习也促进了动态视频的分析, 包括视频预测、视频检索以及动作重定向等任务.

5.2.1 视频预测

视频预测旨在给定一系列上下文帧的情况下预测未来帧, 是一项具有挑战性但十分有趣的任务. Denton 等人提出 DRNET [56], 一种基于自动编码器的模型, 将每一帧分解为不变部分与变化部分, 从而能够一致地生成视频中的未来帧. 视频预测的主要挑战之一在于视觉数据的高维表征空间. 为了解决此问题, Sreekar 等人提出了互信息预测自动编码器 (MIPAE) [160], 将潜在表征分为时间不变 (内容) 与时间变化 (姿态) 两部分, 从而避免对高维视频帧进行直接预测. 他们使用互信息损失与相似度损失促进解耦, 并使用 LSTM 来预测低维姿态表征. 随后, 将内容表征与预测的姿态表征共同解码以生成未来视频帧. Hsieh 等人随后提出 DDPAE [139]

框架, 该方法同样将内容表征与低维姿态表征进行解耦. 他们利用姿态预测网络来根据已有姿态预测未来姿态表征. 基于由预测姿态表征参数化的逆空间 transformer, 内容的不变表征也可以用于预测未来帧.

5.2.2 其他任务

为应对真实场景中大规模模式变化的问题, Kim 等人 [161] 提出一种面向稳健人脸认证的解耦表征学习框架, 通过解耦身份表征与模式表征 (如光照、姿态). 他们首先使用两个编码器分别编码身份与模式表征. 为确保解耦, 他们设计一种排除式策略, 使两个编码器从各自的表征中移除对方的表征. 此外, 他们设计了一种基于重构的策略来加强解耦, 该策略使用解码器通过交换图像对的身份表征来重构原始表征, 然后再重新解耦身份表征和模态表征.

动作重定向旨在将某个源视频中的人体动作迁移到目标视频中. Ma 等人 [162] 指出, 以往的动作重定向方法忽视了传递过程中与主体有关的动作表征, 导致合成结果不自然. 为解决此问题, 他们提出将主体相关动作表征与主体无关动作表征分别解耦, 通过组合源视频的主体无关表征与目标视频的主体相关表征来生成目标动作表征. 为实现这种解耦, 他们为主体相关与主体无关两类表征分别设计三元组损失, 以确保两者的有效分离.

在视频动作识别中, 场景偏置常导致模型依赖背景而非动作本身, 从而在未见过的动作–场景组合上表现欠佳. 为缓解这一问题, Park 等人提出 DEVIAS [163], 通过编码–解码架构显式分离动作与场景表征. 该方法利用 slot attention 提取独立的动作与场景槽, 并通过动作掩码解码器在训练阶段强化动作相关信息的可分离性.

5.3 自然语言处理

解耦表征学习也被广泛探索于自然语言处理任务中, 例如文本生成、风格迁移、语义理解等.

5.3.1 文本表征

最初的解耦表征学习在自然语言处理中的应用主要旨在根据不同准则学习可解耦的文本表征, 通常通过将表征的不同方面编码到不同空间中来实现. He 等人 [23] 在无监督的神经词嵌入模型中引入注意力机制, 以发现具有强识别性、语义连贯的方面, 相较于以往方法显著提升了多方面信息的解耦能力. Bao 等人 [24] 在 VAE 的潜在空间中建模句法信息, 并通过对抗式重建损失对句法与语义空间进行正则化, 从而实现从解耦的句法空间与语义空间生成句子. Cheng 等人 [25] 提出一种带有部分监督的解耦学习框架, 通过优化互信息上界来解耦文本中的风格信息与内容信息. 在保持语义信息的前提下, 该框架在条件文本生成与文本风格迁移任务上表现优异. Wu 等人 [13] 提出一种解耦学习方法, 以提升自然语言处理模型的鲁棒性与泛化能力. Colombo 等人 [164] 通过最小化句子内容与属性的潜在表征之间的互信息来学习可解耦表征, 并基于 KL 散度与 Rényi 散度设计新的变分上界来估计互信息.

5.3.2 风格迁移

在文本风格迁移任务中, 也有大量工作尝试通过解耦表征学习来从文本表征中分离风格信息. Hu 等人 [165] 将 VAE 与属性判别器结合, 以解耦文本数据的内容与属性, 从而生成具有目标情感或时态属性的文本. John 等人 [166] 基于 VAE 引入辅助多任务目标与对抗目标, 用于解耦句子的潜在表征, 在非平行文本风格迁移任务中取得了高性能.

5.3.3 其他任务

在自然语言处理社区中还有一些特定任务中, 解耦表征学习同样发挥着有效作用. Zou 等人 [167] 通过从文本摘要中解耦关键信息与抽象信息来处理语义匹配这一基础任务, 从而学习不同粒度下的内容匹配方式. Dougrez-Lewis 等人 [168] 在对抗学习框架下解耦社交媒体消息的潜在话题, 用于谣言真伪分类. Zhu 等人 [169] 通过弱化句子级信息在风格表征中的影响, 在潜在空间中解耦内容与风格, 以生成具有风格表征的对话回复. 此外, 也有研究 [170, 171] 探索如何将大规模预训练语言模型与解耦表征学习结合. Zhang 等人 [170] 尝试从预训练模型 (如 BERT [172]) 中发现潜在子网络, 从而挖掘可解耦表征, 使其能够分解为输入的不同、互补属性. Zeng 等人 [171] 提出面向任务引导的解耦微调方法, 通过从耦合的表征中解耦任务相关信号来增强表征的泛化能力.

5.4 多模态应用

随着多模态数据的快速发展, 关于多模态任务中的解耦表征学习研究也不断增多, 其中解耦表征学习主要用于不同模态表征的分离、对齐与泛化.

早期工作 [140, 143] 通过鼓励模态特定因素与多模态因素之间的独立性, 研究典型的模态级别解耦. Shi 等人 [141] 提出了多模态生成模型的四项准则, 并基于混合专家层设计了一种多模态 VAE, 从而实现模态间的解耦. Zhang 等人 [144] 提出了一个解耦情感表征对抗网络 (DiSRAN), 用于在跨领域情感分析中减少表达风格的领域偏移. 近期的研究 [142, 145~148] 更倾向于解耦多模态中的丰富信息, 并利用这些信息执行各种下游任务. Alaniz 等人 [142] 提出利用文本的语义结构来解耦视觉数据, 以学习文本与图像之间的统一表征. PPE 框架 [146] 通过利用预训练视觉-语言模型 CLIP [173] 的能力, 实现了基于文本驱动的解耦图像编辑. 类似地, Yu 等人 [145] 通过解耦并利用 CLIP 文本嵌入中的语义, 实现了反事实图像操作. Materzynska 等人 [147] 则将 CLIP 的拼写能力从其视觉概念处理能力中解耦出来.

5.5 推荐系统

解耦表征学习在推荐任务中的应用也受到了广泛关注. 用户行为背后的潜在因素通常复杂且耦合, 在推荐系统中, 解耦后的因素能带来新的视角, 降低复杂度, 并提升效率与可解释性.

推荐中的解耦表征学习主要旨在捕捉用户在不同方面的兴趣. 早期工作 [26, 27, 29, 152] 聚焦于在协同过滤中学习解耦表征. 具体而言, Ma 等人 [26] 提出了 MacridVAE, 用于学习用户在物品上的宏观与微观偏好, 从而支持可控推荐. Wang 等人 [29, 152] 将用户–物品二部图分解为多个解耦子图, 表示不同类型的用户–物品关系. Zhang 等人 [27] 则提出从用户的行为信息与内容信息中同时学习其解耦兴趣. 更多近期研究 [149, 150, 174] 将解耦表征学习应用于序列推荐, 其中用户未来兴趣在解耦意图空间中与其历史行为进行匹配. 此外, 一些研究 [28, 151] 还利用辅助信息提升解耦推荐性能. 特别地, Wang 等人 [28] 同时利用视觉图像与文本描述来提取用户兴趣, 从视觉与文本线索提供推荐解释性. 随后, 他们进一步结合视觉与类别信息, 提供解耦的视觉语义, 从而同时提升推荐结果的可解释性与准确性 [151]. Wang 等人 [175] 进一步在不同环境中学习协同解耦表征, 用于社交推荐.

5.6 图表征学习

随着各类涉及图结构数据的应用不断增长, 图表征学习与图推理方法的需求日益旺盛, 而真实世界的图数据通常包含复杂的关系结构与对象之间的相互依赖 [176, 177]. 因此, 研究者们致力于将解耦表征学习应用于图数据, 在图分析任务中取得了显著进展.

Ma 等人 [30] 指出, 现有方法缺乏对潜在因素复杂纠缠的关注, 并提出了 DisenGCN. 该方法通过邻域路由机制, 根据潜在因素迭代地对邻域进行分割, 从而学习解耦的节点表征. 随后, NED-VAE [153] 被提出作为一种无监督的解耦方法, 可在带属性的图中解耦节点表征与边表征. FactorGCN [154] 则通过将输入图分解为多个因子图来获得图级的解耦表征. 对于每个因子图, 将其分别输入 GNN 模型中, 再将输出聚合以得到解耦的图表征. Li 等人 [155] 提出利用自监督学习来获得解耦的图表征, 并用于分布外泛化 [178]. 给定输入图, 他们提出的 DGCL 方法首先识别图中的潜在因素并得到其因子化表征, 然后进行因子层面的对比学习, 以促使不同的潜在因子能够独立反映其对应的表达性信息. 随后, 他们提出 IDGCL [156], 通过显式强化潜在表征之间的独立性来学习解耦的自监督图表征, 从而提升解耦图表征的质量. Zhang 等人 [179] 则进一步将解耦表征学习应用于增量图神经结构搜索.

5.7 抽象推理

神经网络中的抽象推理最早由 Barrett 等人 [180] 提出. 受到人类智力测试 Raven's Progressive Matrices(RPMs) 的启发, 他们探索了衡量与诱导神经网络抽象推理能力的方法, 并研究了其泛化性能. 特别地, Steenkiste 等人 [181] 研究了解耦表征学习在抽象推理任务中的作用. 他们评估了解耦表征学习模型所获得表征的有效性, 并基于这些解耦的表征训练抽象推理模型. 结果表明, 解耦表征可能带来更好的下游性能, 如在类似 RPMs 的任务中使用更少样本即可更快学习. Locatello 等人 [182] 从观测对中学习解耦表征, 并提出了无需群组标注的自适应基于群组的解耦方法. 他们展示了仅依赖弱监督即可学习解耦表征, 该表征能够体现超越统计相关性的结构, 并在抽象视觉推理任务中表现有效. Amizadeh 等人 [183] 则研究了在视觉问答中推理部分与感知部分的解耦. 他们提出了一种可微的一阶逻辑形式化方法, 使视觉问答中的逻辑推理与视觉感知在回答问题的过程中显式分离, 从而能够独立评估推理与感知能力.

5.8 少样本学习

少样本学习的方法主要分为三类: 基于梯度的方法、基于数据增强的方法以及基于度量学习的方法. 由于解耦表征学习能够获得更加鲁棒且具有更好泛化能力的表征, 它在少样本学习中也得到了一系列应用.

5.8.1 数据增强

处理少样本学习的一个重要途径是数据增强. 针对从图像中提取的类内变化表征往往同时包含类判别表征与与新样本无关的信息的问题, Xu 等人 [184] 提出了一种新的解耦数据增强框架, 该框架能够独立捕获类内变化表征 z_V 与类判别表征 z_I. 他们利用 VAE 从输入图像中提取 z_V, 并通过 $z = z_I + z_V$ 重构图像. 同时, z_V 的分布在所有类别中共享, 其先验为 $N(0, I)$. 表征 z_I 通过最大池化获得, 并利用分类器保证 z_I 能够真正捕获类判别信息. 最终, 通过采样多个 z_V 来实现数据增强.

Lin 等人 [185] 提出一种基于数据增强的少样本图像分类框架, 利用基类的类内变化进行增强. 他们同样将图像解耦为类别特定 (类判别) 表征与外观特定 (类内变化) 表征, 然后通过将新类别样本的类别特定表征与从任意基类样本中提取的多组外观特定表征组合, 生成新类别的附加样本.

5.8.2 度量学习

Cheng 等人 [60] 提出一种基于解耦的度量学习方法 DFR, 用于少样本图像分类. 他们使用两个编码器 E_{cls} 与 E_{var} 分别提取类别特定表征与类别无关表征 (如背景、图像风格). 其中, 类别特定表征也称为判别表征, 用

于最终的少样本分类; 而类别无关表征对分类是冗余的. 为了实现解耦, 他们通过梯度反转层 [186] 以及类别判别器来最小化 E_{var} 捕获的类判别信息. 此外, 重构损失和分类损失也进一步促进了解耦.

Tokmakov 等人 [187] 提出一种解耦表征框架用于少样本图像分类. 他们收集了类别级别的属性标注, 如物体部件 (例如翅膀颜色) 和场景元素 (例如草地). 随后, 他们将图像表征解耦为对应不同属性的部分, 并假设图像表征等于所有属性表征的总和, 使用一定的距离损失函数来约束该结构. 一个正交损失被用于进一步促进解耦. 实验结果显示, 该解耦表征在样本极少的新类别上具有更好的泛化能力.

Prabhudesai 等人 [188] 提出了解耦 3D 原型网络 D3DP-Nets, 用于少样本概念学习. 他们首先通过训练编码器-解码器, 将 RGB-D 图像编码为 3D 表征图, 然后通过两个编码器将 3D 表征图解耦为 3D 形状编码与 1D 风格编码. 基于形状编码和风格编码, 模型能够计算每个概念的原型. 由于具备解耦特性, 分类器可以仅关注每个概念的关键表征, 从而在少样本条件下具备更强的泛化能力.

5.8.3 其他

Sun 等人 [189] 提出一种解耦框架 SAVAE, 用于少样本字体风格生成. 他们将字符输入解耦为内容表征与风格表征, 使用风格推断网络编码风格码, 使用内容识别网络提取内容信息. 在训练过程中, 他们设计了交叉成对优化算法, 以确保风格编码器能够纯粹捕获风格信息, 也就是说具有相同风格的不同字符应当具有相同的风格码. 通过推断风格表征, 该模型在未见风格上的少样本泛化能力非常强.

Guo 等人 [190] 提出一种用于少样本视觉问答的解耦框架. 他们工作的关键洞察是: 将新类别的对象拆分为模型已学习过的部分, 可以促进新类别对象的学习. 因此, 他们设计了一个属性网络用于解耦答案的属性. 在少样本阶段, 网络通过最小化答案与其属性之和之间的距离来约束答案. 虽然新答案的训练样本很少, 但其属性已在基础阶段学习, 因此该框架在少样本设定下具有良好的泛化能力.

5.9 分布外问题

解耦表征学习具备学习更具鲁棒性的表征的潜力, 这些表征能够泛化至分布外场景, 因为它能够将与任务目标相关的关键表征与无关因素进行分离. Li 等人 [157] 发现, 学习解耦的图表征能够提升图神经网络的分布外泛化能力. 其提出的 OOD-GNN 模型通过迭代优化样本图的权重和图编码器, 消除输出表征各维度之间的统计依赖, 从而促进图表征的解耦. 涵盖真实而具有挑战性的图分布偏移的实验表明, 该解耦的 OOD-GNN 模型相较多种代表性的最新图神经网络方法取得了显著的分布外泛化提升.

Dittadi 等人 [191] 提出一个框架, 用于评估解耦表征在下游任务中的分布外泛化能力. 他们设计了因子回归任务, 并构建了多种分布外设置. 结果表明, 在某些情况下, 解耦表征能够提升分布外性能.

Yoo 等人 [192] 为有向网络嵌入中与度相关的分布偏移问题提出了解耦框架, 以增强其分布外泛化能力. 他们为每条有向边分配六种不同的因子, 并为每个节点学习相应的解耦嵌入. 实验结果表明, 该解耦嵌入在不同程度的度分布偏移下均显著促进了分布外泛化.

Zhang 等人 [193] 提出一个用于领域泛化的解耦表征学习框架. 他们将领域不变因子与领域特异因子分离到不同的子空间中, 仅使用领域不变表征进行预测. 实验结果显示, 该解耦表征在多个数据集上显著促进了分布外泛化.

Mu 等人 [194] 也提出一个用于关节化形状表征的解耦表征学习框架. 他们通过分离形状表征与关节表征来实现解耦. 基于这些解耦表征, 模型能够为未见过的实例生成具有未见过关节角度的形状, 从而展示了在分布外场景中的强泛化能力.

6 解耦表征学习在不同任务中的设计

在本节中, 我们讨论解耦表征学习在实际应用中的常用策略, 为针对具体任务设计不同的解耦表征学习模型提供启发. 我们总结了解耦表征学习模型设计的两个关键方面:i) 根据具体任务设计合适的表征结构;ii) 设计相应的损失函数, 使表征具有解耦性但又不丢失任务相关信息.

6.1 表征结构的设计

给定一个具体任务, 我们首先需要确定解耦表征的结构. 为此, 我们需要考虑潜在生成因子的结构. 具体而言, 一方面, 需要考虑这些因子是具有层次结构还是平坦结构. 如果是前者, 则应采用层次化的解耦表征学习方法; 否则应使用平坦式的解耦表征学习方法. 另一方面, 需要考虑需要解耦的生成因子的数量和粒度. 如在第 3.2.1 节中讨论的那样, 在简单场景中, 适合使用针对多个细粒度因子的维度级解耦方法, 而在更复杂和真实世界的场景中, 则适合使用针对少量粗粒度因子的向量级解耦方法.

本节基于“维度级或向量级的方法”这一分类讨论解耦表征结构的设计. 考虑两种互补的结构:i) 维度级: 使用一个整体向量表征 z, 其具有细粒度结构;ii) 向量级: 使用两个或更多独立向量 $z_1, z_2, \ldots$ 来表征数据表征的不同部分, 其具有粗粒度结构. 为保证解耦性, 方法 i) 通常要求 z 在维度上相互独立, 而方法 ii) 通常要求 z_i 与 $z_j (i \neq j)$ 相互独立.

如果在应用中选择维度级的方法, 则可选的典型模型包括各种基于 VAE 和 GAN 的方法, 已在第 3 节详细介绍. 在此情况下, 我们可以使用 VAE 或 GAN 作为骨干网络, 并设计额外的损失函数以适配具体任务. 我们也可以采用其他模型结构, 例如 InfoSwap [61], 其使用多层编码器提取任务相关表征, 并基于信息瓶颈逐层压缩表征以丢弃任务无关信息.

对于向量级的方法, 通常有两类获得多个潜在向量的方式:i) 预设这些潜在向量;ii) 使用不同的编码器以原始表征为输入, 将原始整体向量拆分为多个不同向量. 例如, DR-GAN [57] 显式设置一个潜在向量表征姿态, 并使用编码器从输入图像中提取身份编码, 并通过监督损失函数确保姿态编码与身份编码能够分别捕获姿态与身份信息. Liu 等人 [55] 使用两个编码器 (运动编码器与外观编码器) 分别通过原始表征提取运动表征与外观表征. Cheng 等人 [60] 使用两个编码器 E_{cls} 和 E_{var} 分别提取类别相关表征和类别无关表征. DRNET [56] 也使用两个编码器分别提取姿态表征与内容表征. DRANet [21] 则仅使用一个编码器提取内容表征, 并通过从原始表征中减去内容表征获得风格表征. 类似地, Wu 等人 [195] 使用一个编码器从图像表征图中提取领域不变表征, 然后通过减去领域不变表征获得领域特定表征.

我们必须指出, 无论选择何种模型结构, 都必须设计合适的损失函数, 以确保表征能够解耦, 同时不丢失数据中所携带的信息.

6.2 损失函数的设计

在本节中, 我们将根据不同的模型类型 (即生成模型与判别模型) 讨论用于保证解耦性与信息量的损失函数设计. 总体而言, 我们将损失函数总结为

$$L = \lambda_1 L_{\text{re}} + \lambda_2 L_{\text{disen}} + \lambda_3 L_{\text{task}},$$

其中 L_{re} 表示重建损失, L_{disen} 表示解耦损失, L_{task} 表示特定任务损失.

重建损失对于生成任务来说通常是必需的, 它确保解耦表征在语义上是有意义的, 并且可以恢复原始数据. 解耦损失用于强化表征的解耦性. 此外, 重建损失有时也有助于解耦, 因为它要求模型必须借助这些解耦表征正确地重建数据. 任务损失与具体任务目标直接相关. 在联合优化任务损失与解耦模块的情况下, 任务损失通

常可以为解耦过程提供指导. 相反, 如果采用两阶段方案, 即先训练解耦模块, 再将解耦表征用于下游任务, 则任务损失无法指导解耦过程.

6.2.1 生成任务

第 3 节中提到的各种基于 VAE 的模型都在 ELBO 中包含显式的重建损失, 并使用额外的正则项作为解耦损失. 对于基于 GAN 的方法, 对抗损失也可视作重建损失, 而解耦损失可以是互信息约束, 如 InfoGAN [9] 和 IB-GAN [47] 所采用的那样.

DRANet [21] 采用 L_1 损失作为重建损失, 并使用对抗损失作为任务损失, 以确保任务目标 (即跨域图像适配) 得以实现. 该任务损失根据基于解耦表征生成的图像计算, 因此也鼓励模型获得正确的解耦表征. 此外, 它还使用一致性损失与感知损失以增强解耦效果.

DRNET [56] 采用 L_2 重建损失, 并使用相似性损失与对抗损失共同保证解耦性. 由于它是一个两阶段方案 (解耦表征旨在用于视频预测等下游任务), 因此忽略任务损失.

InfoSwap [61] 使用基于信息瓶颈理论的信息压缩损失作为解耦损失, 同时使用对抗损失作为人脸身份交换的任务损失, 并使用一致性损失作为重建损失.

MAP-IVR [55] 除了使用 L_2 重建损失以分别保证运动表征与外观表征捕获动态与静态信息之外, 还引入余弦相似度损失以强化运动与外观表征之间的正交性. MAP-IVR 同样没有任务损失, 因为它属于两阶段方案, 学习到的运动与外观表征用于下游任务, 例如动作图像到视频检索.

此外, 一些方法无需显式解耦损失函数, 而通过额外监督来强制解耦, 例如 DR-GAN [57] 与 DNAGAN [59].

6.2.2 判别任务

判别任务有时也需要重建损失. 例如, 判别任务可以使用 VAE 作为表征提取的骨干网络, 或者利用重建损失以确保解耦表征能够正确恢复数据. 判别任务通常不限制特定的骨干模型, 它们使用由 VAE 或 GAN 等适当模型编码的潜在解耦表征, 并在此基础上加入目标任务所需的任务损失 (如图像分类、推荐、神经架构搜索等).

例如, Hamaguchi 等人 [136] 在采用两个 VAE 对图像对进行编码的基础上加入相似性损失与激活损失, 旨在使公共表征编码输入图像对中的不变因素. 这两个损失鼓励模型分别学习图像的公共表征和特定表征. 该方法使用分类损失作为任务损失以实现稀有事件检测, 同时也指导解耦编码器的训练.

Wu 等人 [195] 使用正交损失促进领域不变表征与领域特定表征之间的独立性, 并采用对抗机制鼓励领域特定表征捕获更多领域特定信息. 其任务损失为用于领域自适应目标检测的检测损失.

Cheng 等人 [60] 使用梯度反转层与类别判别损失以最小化类别无关编码器中包含的类别特定信息. 他们使用 L_1 重建损失与翻译损失以确保解耦表征可以正确重建数据, 并使用分类损失完成小样本图像分类.

表 6 总结了上述生成任务与判别任务的损失函数设计.

7 未来方向

最后, 我们通过指出一些值得未来继续探索的潜在研究方向来总结本文.

构建更完善的理论基础与探索前沿方向. 尽管解耦表征学习在经验上已被证明是有效的, 但目前仍缺乏严格的数学保证来说明特定表征是否总能被完全解耦, 或最多能解耦到什么程度. 此外, 解耦性与泛化能力、鲁

表 6 若干生成任务和判别任务中损失函数的总结.
The summary of loss functions of several generative and discriminative tasks.

Methods	reconstruction loss	disentanglement loss	task loss
VAE-based Approaches	VAE loss	Extra regularizers added to ELBO	–
InfoGAN [9]	GAN loss	Maximizing $\lambda\mathcal{I}(c; G(z, c))$	–
DRANet [21]	L1 loss	Perceptual loss,consistency loss	Adversarial loss
DRNET [56]	L2 loss	Similarity loss,adversarial loss	–
InfoSwap [61]	Consistency loss	Information-compression loss	Adversarial loss
MAP-IVR [55]	L2 loss	$\mathcal{L}_{\text{orth}} = \cos(m^0, a^0)$, $\mathcal{L}_{\text{class}}$	–
Hamaguchi et al. [136]	VAE loss	Similarity loss, activation loss	Classification loss
Cheng et al. [60]	L1 loss	Discriminative loss	Classification loss
Wu et al. [195]	–	Orthogonality loss, adversarial loss	Detection loss

棒性之间的理论联系仍有待进一步发掘. 为了指导未来的科研工作, 以下几个理论方向可能是极具潜力的研究起点:

1. **探索超越 "统计独立性" 的数学与物理先验**. 传统的解耦往往依赖于特征间的严格统计独立, 这在真实复杂场景中难以成立. 未来的研究可以放宽这一假设, 转而从自然界的生成规律中寻找新的归纳偏置. 例如, 探索基于物理对称性、空间几何变换或独立因果机制 (IMA) 的全新数学框架. 这类方法有助于在更符合现实物理规律的设定下, 重新定义和保证解耦表征的可识别性.

2. **融合结构因果模型以实现真实的机理解耦**. 真实世界中的概念大多存在因果关联而非完全孤立. 下一步工作可以考虑将因果推断引入表征学习, 研究如何在低维潜空间中恢复具有层级或网络结构的因果变量. 探索如何利用弱监督信号、干预数据或时序动态来精确定位因果机制, 将是推动模型从 "相关性拟合" 走向 "真实因果推理" 的核心基础.

3. **建立解耦与分布外泛化及鲁棒性的理论桥梁**. 未来的研究应当致力于建立统一的理论框架, 从理论上严格解释解耦表征如何转化为下游任务的优势. 例如, 通过信息论或泛化边界等理论工具, 论证解耦特征如何有效地过滤虚假相关性, 并抵御对抗性扰动. 揭示解耦表示在分布偏移下降低样本复杂度的内在机理, 将为开发高可靠性、高泛化性的基础模型提供坚实的理论支撑.

改进评估指标.

构建更客观、标准化、全面且普适的基准与指标, 用于评估解耦表征的质量与有效性, 是非常重要的. 这将有助于公平、系统地比较不同方法, 并推动未来的研究进一步发展.

将可解释性提升到新的层次.

解耦表征学习的主要动机之一是学习可解释的数据表征. 然而, 遗憾的是, 现有方法都无法完全解释潜在空间中向量的语义含义. 因此, 我们认为持续提升解耦表征的可解释性仍然非常重要. 相关方向包括:

1. 人类可理解的表征. 解耦表征的最终目标应不仅是可分解的, 而且应是符合人类直觉、易于理解的, 从而促进可解释人工智能的发展. 这包括将学习到的因子与人类语义概念对齐.

2. 交互式解耦表征学习. 探索支持人类参与的交互式学习环境, 使人类能够引导解耦过程, 确保模型学习的表征与用户定义的关键概念相一致.

解耦基础模型.

基础模型 (如 ChatGPT、Stable Diffusion 等) 目前在研究领域中占据主导地位, 在各种下游任务中表现强大. 这些模型的力量主要来自大规模训练数据与数十亿级的模型参数. 我们相信基础模型可以从解耦表征学习中受益, 主要体现在以下方面:

1. 面向特定任务的高效适配. 对于特定下游任务, 基础模型中通常存在冗余. 现有方法通常采用迁移学习技术 (如微调) 进行适配, 但迁移学习无法区分预训练模型中哪些知识与目标任务相关, 可能导致冗余. 解耦学习可以将任务相关与任务无关部分分离, 从而使适配更精确、更高效.

2. 提升基础模型的可解释性. 尽管基础模型具有强大能力甚至一定的推理能力, 但其仍然是缺乏解释性的黑盒模型. 解耦表征学习具有提升透明性与可解释性的潜力, 例如识别模型不同模块中表征的作用, 并将推理过程分解为人类可理解的独立成分.

探索解耦表征学习在复杂真实场景中的潜力.

如前所述, 解耦表征学习的理论研究主要集中在简单数据集上, 这可能无法跟上各种复杂应用 (如视觉问答、文本到图像生成、文本到视频生成) 和复杂模型 (如扩散模型) 在真实世界中的需求.

我们认为有必要进一步探索解耦表征学习在高级真实场景中的能力, 方向包括:

1. 设计适用于真实世界的更有效架构. 真实场景通常涉及粗粒度且复杂的生成因子, 因此需要更有效的结构. 例如, 胶囊网络可能是一个有前景的方向, 而传统的按维度解耦 (如基于 VAE 的方法) 可能不适用于真实复杂场景.

2. 探索 DRL 在高级任务与模型中的应用潜力. 例如, 在扩散模型的文本到图像生成中, 解耦带来的可因子化性、鲁棒性与泛化能力可能为这些任务带来提升.

3. 跨学科应用场景下的新需求. 现实应用需求跨越诸多领域, 如医疗、自动驾驶与金融. 理解这些领域的特定需求可以启发解耦表征学习的新方法.

关注伦理与公平的人工智能.

最后, 我们应始终牢记, 通过利用解耦表征来识别与消除偏见的来源, 以最大程度地减少 AI 系统中的伦理问题与不公平性.

参考文献

1 Geirhos R, Jacobsen J H, Michaelis C, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2020, 2: 665–673
2 Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 2013, 35: 1798–1828
3 Lake B M, Ullman T D, Tenenbaum J B, et al. Building machines that learn and think like people. Behavioral and brain sciences, 2017, 40
4 Burgess C, Kim H. 3d shapes dataset. https://github.com/deepmind/3d-shapes, 2018
5 Chen R T, Li X, Grosse R, et al. Isolating sources of disentanglement in vaes. In: Proceedings of Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2019. 2615–2625
6 Higgins I, Matthey L, Pal A, et al. beta-vae: Learning basic visual concepts with a constrained variational framework. In: Proceedings of International Conference on Learning Representations (ICLR), 2016
7 Kim H, Mnih A. Disentangling by factorising. In: Proceedings of International Conference on Machine Learning. PMLR, 2018. 2649–2658
8 Lee H Y, Tseng H Y, Huang J B, et al. Diverse image-to-image translation via disentangled representations. In: Proceedings of Proceedings of the European Conference on Computer Vision (ECCV), 2018
9 Chen X, Duan Y, Houthooft R, et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. 2180–2188

10 Larsen A B L, Sønderby S K, Larochelle H, et al. Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of International Conference on Machine Learning. PMLR, 2016. 1558–1566

11 Yang M, Liu F, Chen Z, et al. Causalvae: Disentangled representation learning via neural structural causal models. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 9593–9602

12 Ghandeharioun A, Kim B, Li C L, et al. Dissect: Disentangled simultaneous explanations via concept traversals. arXiv preprint arXiv:2105.15164, 2021

13 Wu J, Li X, Ao X, et al. Improving robustness and generality of nlp models using disentangled representations. arXiv preprint arXiv:2009.09587, 2020

14 Suter R, Miladinovic D, Schölkopf B, et al. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In: Proceedings of International Conference on Machine Learning. PMLR, 2019. 6056–6065

15 Lee J, Kim E, Lee J, et al. Learning debiased representation via disentangled feature augmentation. Advances in Neural Information Processing Systems, 2021, 34: 25123–25133

16 Kingma D P, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013

17 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems. Curran Associates, Inc., 2014

18 Higgins I, Amos D, Pfau D, et al. Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230, 2018

19 Zhu J Y, Zhang Z, Zhang C, et al. Visual object networks: Image generation with disentangled 3d representations. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 118–129

20 Gonzalez-Garcia A, van de Weijer J, Bengio Y. Image-to-image translation for cross-domain disentanglement. In: Proceedings of Advances in Neural Information Processing Systems. Curran Associates, Inc., 2018

21 Lee S, Cho S, Im S. Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 15252–15261

22 Liu Y, Sangineto E, Chen Y, et al. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 10785–10794

23 He R, Lee W S, Ng H T, et al. An unsupervised neural attention model for aspect extraction. In: Proceedings of Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017. 388–397

24 Bao Y, Zhou H, Huang S, et al. Generating sentences from disentangled syntactic and semantic spaces. In: Proceedings of Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, 2019. 6008–6019

25 Cheng P, Min M R, Shen D, et al. Improving disentangled text representation learning with information-theoretic guidance. In: Proceedings of Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 7530–7541

26 Ma J, Zhou C, Cui P, et al. Learning disentangled representations for recommendation. In: Proceedings of Advances in Neural Information Processing Systems. Curran Associates, Inc., 2019

27 Zhang Y, Zhu Z, He Y, et al. Content-collaborative disentanglement representation learning for enhanced recommendation. In: Proceedings of Fourteenth ACM Conference on Recommender Systems, 2020. 43–52

28 Wang X, Chen H, Zhu W. Multimodal disentangled representation for recommendation. In: Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021. 1–6

29 Wang X, Jin H, Zhang A, et al. Disentangled graph collaborative filtering. In: Proceedings of Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020. 1001–1010

30 Ma J, Cui P, Kuang K, et al. Disentangled graph convolutional networks. In: Proceedings of International Conference on Machine Learning. PMLR, 2019. 4212–4221

31 Liu X, Sanchez P, Thermos S, et al. Learning disentangled representations in the imaging domain. Medical Image Analysis, 2022, page 102516

32 Dupont E. Learning disentangled joint continuous and discrete representations. In: Proceedings of Advances in Neural Information Processing Systems. Curran Associates, Inc., 2018

33 Kim M, Wang Y, Sahu P, et al. Relevance factor vae: Learning and identifying disentangled factors. arXiv preprint arXiv:1902.01568, 2019

34 Burgess C P, Higgins I, Pal A, et al. Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599, 2018

35 Kumar A, Sattigeri P, Balakrishnan A. Variational inference of disentangled latent concepts from unlabeled observations. In: Proceedings of International Conference on Learning Representations, 2018

36 Bouchacourt D, Tomioka R, Nowozin S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence, 2018

37 Bing S, Fortuin V, Rätsch G. On disentanglement in gaussian process variational autoencoders. arXiv preprint arXiv:2102.05507, 2021

38 Caselles-Dupré H, Garcia Ortiz M, Filliat D. Symmetry-based disentangled representation learning requires interaction with environments. Advances in Neural Information Processing Systems, 2019, 32

39 Quessard R, Barrett T, Clements W. Learning disentangled representations and group structure of dynamical environments. Advances

in Neural Information Processing Systems, 2020, 33: 19727–19737
40 Yang T, Ren X, Wang Y, et al. Towards building a group-based unsupervised representation disentanglement framework. In: Proceedings of International Conference on Learning Representations, 2022
41 Wang T, Yue Z, Huang J, et al. Self-supervised learning disentangled group representation as feature. Advances in Neural Information Processing Systems, 2021, 34
42 Pearl J. Causality. Cambridge university press, 2009
43 Shen X, Liu F, Dong H, et al. Disentangled generative causal representation learning. arXiv preprint arXiv:2010.02637, 2020
44 Li Y, Mandt S. Disentangled sequential autoencoder. arXiv preprint arXiv:1803.02991, 2018
45 Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019
46 Zhu X, Xu C, Tao D. Commutative lie group vae for disentanglement learning. In: Proceedings of International Conference on Machine Learning. PMLR, 2021. 12924–12934
47 Jeon I, Lee W, Pyeon M, et al. Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks. In: Proceedings of 35th AAAI Conference on Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence. ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE, 2021. 7926–7934
48 Lin Z, Thekumparampil K K, Fanti G C, et al. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. 2019
49 Zhu X, Xu C, Tao D. Where and what? examining interpretable disentangled representations. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 5861–5870
50 Wei Y, Shi Y, Liu X, et al. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 6721–6730
51 Yang T, Wang Y, Lv Y, et al. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. arXiv preprint arXiv:2301.13721, 2023
52 Chen H, Zhang Y, Wang X, et al. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374, 2023
53 Chen H, Zhang Y, Wang X, et al. Disendreamer: Subject-driven text-to-image generation with sample-aware disentangled tuning. IEEE Transactions on Circuits and Systems for Video Technology, 2024
54 Kwon M, Jeong J, Uh Y. Diffusion models already have a semantic latent space. In: Proceedings of The Eleventh International Conference on Learning Representations, 2022
55 Liu L, Li J, Niu L, et al. Activity image-to-video retrieval by disentangling appearance and motion. In: Proceedings of Proc. AAAI, 2021. 1–9
56 Denton E, Birodkar V. Unsupervised learning of disentangled representations from video. In: Proceedings of Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 4417–4426
57 Tran L, Yin X, Liu X. Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017. 1415–1424
58 Singh K K, Ojha U, Lee Y J. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In: Proceedings of Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019. 6490–6499
59 Xiao T, Hong J, Ma J. Dna-gan: Learning disentangled representations from multi-attribute images. arXiv preprint arXiv:1711.05415, 2017
60 Cheng H, Wang Y, Li H, et al. Disentangled feature representation for few-shot image classification. arXiv preprint arXiv:2109.12548, 2021
61 Gao G, Huang H, Fu C, et al. Information bottleneck disentanglement for identity swapping. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3404–3413
62 Ross A, Doshi-Velez F. Benchmarks, algorithms, and metrics for hierarchical disentanglement. In: Proceedings of International Conference on Machine Learning. PMLR, 2021. 9084–9094
63 Li Z, Murkute J V, Gyawali P K, et al. Progressive learning and disentanglement of hierarchical representations. arXiv preprint arXiv:2002.10549, 2020
64 Tong B, Wang C, Klinkigt M, et al. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 11467–11476
65 Li X, Zhang S, Hu J, et al. Image-to-image translation via hierarchical style disentanglement. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 8639–8648
66 Hjelm R D, Fedorov A, Lavoie-Marchildon S, et al. Learning deep representations by mutual information estimation and maximization. In: Proceedings of International Conference on Learning Representations (ICLR), 2018
67 Alemi A A, Fischer I, Dillon J V, et al. Deep variational information bottleneck. In: Proceedings of International Conference on Learning Representations (ICLR), 2017
68 Locatello F, Bauer S, Lucic M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations. In:

Proceedings of international conference on machine learning. PMLR, 2019. 4114–4124
69 Locatello F, Tschannen M, Bauer S, et al. Disentangling factors of variations using few labels. In: Proceedings of International Conference on Learning Representations, 2020
70 Kulkarni T D, Whitney W, Kohli P, et al. Deep convolutional inverse graphics network. arXiv preprint arXiv:1503.03167, 2015
71 Bengio Y, Deleu T, Rahaman N, et al. A meta-transfer objective for learning to disentangle causal mechanisms. In: Proceedings of International Conference on Learning Representations, 2020
72 Khemakhem I, Kingma D, Monti R, et al. Variational autoencoders and nonlinear ica: A unifying framework. In: Proceedings of International Conference on Artificial Intelligence and Statistics. PMLR, 2020. 2207–2217
73 Liu T, Li Q, et al. Disentangled representation learning with the gromov-monge gap. In: Proceedings of ICLR, 2025
74 Lin Z, Lu J, Shen W, et al. Disentangled clothed avatar generation from text. In: Proceedings of ECCV, 2024
75 Zhang B, Shi H, Wu T, et al. Fluxspace: Disentangled semantic editing in rectified flow models. In: Proceedings of CVPR, 2025
76 Gong R, Tang S, Geiger A. Can diffusion models disentangle? In: Proceedings of NeurIPS, 2025
77 Mu Z, Yang X, Sun S, et al. Self-supervised disentangled representation learning for robust target speech extraction, 2024
78 Wu A, Zheng W S. Factorized diffusion autoencoder for unsupervised disentangled representation learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, URL https://ojs.aaai.org/index.php/AAAI/article/view/28407
79 Qi T, Fang S, Wu Y, et al. Deadiff: An efficient stylization diffusion model with disentangled representations, 2024
80 Ding G, Yang C, Wang S, et al. Dis²booth: Learning image distribution with disentangled features for text-to-image diffusion models. 2025, URL https://ojs.aaai.org/index.php/AAAI/article/view/32279
81 LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86: 2278–2324
82 Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild. In: Proceedings of Proceedings of the IEEE international conference on computer vision, 2015. 3730–3738
83 Aubry M, Maturana D, Efros A, et al. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In: Proceedings of CVPR, 2014
84 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10684–10695
85 Kawar B, Zada S, Lang O, et al. Imagic: Text-based real image editing with diffusion models. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 6007–6017
86 Wu J Z, Ge Y, Wang X, et al. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 7623–7633
87 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020, 33: 6840–6851
88 Liu L, Ren Y, Lin Z, et al. Pseudo numerical methods for diffusion models on manifolds. In: Proceedings of International Conference on Learning Representations, 2021
89 Lu C, Zhou Y, Bao F, et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems, 2022, 35: 5775–5787
90 Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 2021, 34: 8780–8794
91 Chen H, Wang X, Zeng G, et al. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. arXiv preprint arXiv:2311.00990, 2023
92 Shen Y, Zhou B. Closed-form factorization of latent semantics in gans. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 1532–1540
93 Khrulkov V, Mirvakhabova L, Oseledets I, et al. Disentangled representations from non-disentangled models. arXiv preprint arXiv:2102.06204, 2021
94 Voynov A, Babenko A. Unsupervised discovery of interpretable directions in the gan latent space. In: Proceedings of International conference on machine learning. PMLR, 2020. 9786–9796
95 Ren X, Yang T, Wang Y, et al. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In: Proceedings of International Conference on Learning Representations, 2021
96 Reed S E, Zhang Y, Zhang Y, et al. Deep visual analogy-making. In: Proceedings of Cortes C, Lawrence N, Lee D, et al., editors, Advances in Neural Information Processing Systems. Curran Associates, Inc., 2015
97 Gondal M W, Wüthrich M, Đorđe Miladinović, et al. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset, 2019
98 Chen R T Q, Li X, Grosse R, et al. Isolating sources of disentanglement in variational autoencoders, 2019
99 Estermann B, Wattenhofer R. Dava: Disentangling adversarial variational autoencoder, 2023
100 Lin Z, Thekumparampil K K, Fanti G, et al. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans, 2020

101 Härkönen E, Hertzmann A, Lehtinen J, et al. Ganspace: Discovering interpretable gan controls, 2020

102 Yang T, Lan C, Lu Y, et al. Diffusion model with cross attention as an inductive bias for disentanglement, 2024

103 Jun Y, Park J, Choo K, et al. Disentangling disentangled representations: Towards improved latent units via diffusion models, 2024

104 Eastwood C, Williams C K. A framework for the quantitative evaluation of disentangled representations. In: Proceedings of International Conference on Learning Representations, 2018

105 Yeats E, Liu F, Womble D, et al. Nashae: Disentangling representations through adversarial covariance minimization, 2022

106 Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018

107 Zhao S, Song J, Ermon S. Infovae: Information maximizing variational autoencoders, 2018

108 Preechakul K, Chatthee N, Wizadwongsa S, et al. Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 10619–10629

109 Wang Y, Schiff Y, Gokaslan A, et al. Infodiffusion: Representation learning using information maximizing diffusion models, 2023

110 Xiang S, Gu Y, Xiang P, et al. Disunknown: Distilling unknown factors for disentanglement learning. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 14810–14819

111 Reddy A G, Benin Godfrey L, Balasubramanian V N. On causally disentangled representations. arXiv preprint arXiv:2112.05746, 2021

112 Greenland S, Pearl J, Robins J M. Confounding and collapsibility in causal inference. Statistical science, 1999, 14: 29–46

113 Yu Y, Chen J, Gao T, et al. Dag-gnn: Dag structure learning with graph neural networks. In: Proceedings of International Conference on Machine Learning. PMLR, 2019. 7154–7163

114 Ribeiro F D S, Duarte K, Everett M, et al. Learning with capsules: A survey. arXiv preprint arXiv:2206.02664, 2022

115 Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules. Advances in neural information processing systems, 2017, 30

116 Patrick M K, Adekoya A F, Mighty A A, et al. Capsule networks–a survey. Journal of King Saud University-computer and information sciences, 2022, 34: 1295–1310

117 Mazzia V, Salvetti F, Chiaberge M. Efficient-capsnet: Capsule network with self-attention routing. Scientific reports, 2021, 11: 14634

118 Hu M f, Liu J w. β-capsnet: learning disentangled representation for capsnet by information bottleneck. Neural Computing and Applications, 2023, 35: 2503–2525

119 Locatello F, Weissenborn D, Unterthiner T, et al. Object-centric learning with slot attention. Advances in Neural Information Processing Systems, 2020, 33: 11525–11538

120 Seitzer M, Horn M, Zadaianchuk A, et al. Bridging the gap to real-world object-centric learning. In: Proceedings of The Eleventh International Conference on Learning Representations, 2022

121 Singh G, Wu Y F, Ahn S. Simple unsupervised object-centric learning for complex and naturalistic videos. Advances in Neural Information Processing Systems, 2022, 35: 18181–18196

122 Elsayed G, Mahendran A, van Steenkiste S, et al. Savi++: Towards end-to-end object-centric learning from real-world videos. Advances in Neural Information Processing Systems, 2022, 35: 28940–28954

123 Sylvain T, Zhang P, Bengio Y, et al. Object-centric image generation from layouts. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 2647–2655

124 Cheng B, Girshick R, Dollár P, et al. Boundary iou: Improving object-centric image segmentation evaluation. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 15334–15342

125 Wu Z, Dvornik N, Greff K, et al. Slotformer: Unsupervised visual dynamics simulation with object-centric models. arXiv preprint arXiv:2210.05861, 2022

126 Ferraro S, Van de Maele T, Mazzaglia P, et al. Disentangling shape and pose for object-centric deep active inference models. In: Proceedings of International Workshop on Active Inference. Springer, 2022. 32–49

127 Li N, Eastwood C, Fisher R. Learning object-centric representations of multi-object scenes from multiple views. Advances in Neural Information Processing Systems, 2020, 33: 5656–5666

128 Zaidi J, Boilard J, Gagnon G, et al. Measuring disentanglement: A review of metrics. arXiv preprint arXiv:2012.09276, 2020

129 Eastwood C, Williams C K. A framework for the quantitative evaluation of disentangled representations. In: Proceedings of International Conference on Learning Representations, 2018

130 Ridgeway K, Mozer M C. Learning deep disentangled embeddings with the f-statistic loss. Advances in neural information processing systems, 2018, 31

131 Tokui S, Sato I. Disentanglement analysis with partial information decomposition. In: Proceedings of International Conference on Learning Representations, 2022

132 Do K, Tran T. Theory and evaluation metrics for learning disentangled representations. arXiv preprint arXiv:1908.09961, 2019

133 Wu Z, Lischinski D, Shechtman E. Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 12863–12872

134 Träuble F, Creager E, Kilbertus N, et al. On disentangled representations learned from correlated data. In: Proceedings of International Conference on Machine Learning. PMLR, 2021. 10401–10412

135 Zeng X, Pan Y, Wang M, et al. Realistic face reenactment via self-supervised disentangling of identity and pose. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence, 2020. 12757–12764

136 Hamaguchi R, Sakurada K, Nakamura R. Rare event detection using disentangled representation learning. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

137 Sanchez E H, Serrurier M, Ortner M. Learning disentangled representations via mutual information estimation. In: Proceedings of European Conference on Computer Vision. Springer, 2020. 205–221

138 Feng Z, Xu C, Tao D. Self-supervised representation learning by rotation feature decoupling. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

139 Hsieh J T, Liu B, Huang D A, et al. Learning to decompose and disentangle representations for video prediction. Advances in neural information processing systems, 2018, 31

140 Hsu W N, Glass J. Disentangling by partitioning: A representation learning framework for multimodal sensory data. arXiv preprint arXiv:1805.11264, 2018

141 Shi Y, Paige B, Torr P, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. Advances in Neural Information Processing Systems, 2019, 32

142 Alaniz S, Federici M, Akata Z. Compositional mixture representations for vision and text. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022. 4202-4211

143 Tsai Y H H, Liang P P, Zadeh A, et al. Learning factorized multimodal representations. In: Proceedings of International Conference on Learning Representations, 2018

144 Zhang Y, Zhang Y, Guo W, et al. Learning disentangled representation for multimodal cross-domain sentiment analysis. IEEE Transactions on Neural Networks and Learning Systems, 2022

145 Yu Y, Zhan F, Wu R, et al. Towards counterfactual image manipulation via clip. In: Proceedings of Proceedings of the 30th ACM International Conference on Multimedia, 2022. 3637–3645

146 Xu Z, Lin T, Tang H, et al. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 18229–18238

147 Materzyńska J, Torralba A, Bau D. Disentangling visual and written concepts in clip. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 16410–16419

148 Zou W, Ding J, Wang C. Utilizing bert intermediate layers for multimodal sentiment analysis. In: Proceedings of 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022. 1–6

149 Ma J, Zhou C, Yang H, et al. Disentangled self-supervision in sequential recommenders. In: Proceedings of KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020. ACM, 2020. 483–491

150 Chen H, Chen Y, Wang X, et al. Curriculum disentangled recommendation with noisy multi-feedback. Advances in Neural Information Processing Systems, 2021, 34: 26924–26936

151 Wang X, Chen H, Zhou Y, et al. Disentangled representation learning for recommendation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022

152 Hu L, Xu S, Li C, et al. Graph neural news recommendation with unsupervised preference disentanglement. In: Proceedings of Proceedings of the 58th annual meeting of the association for computational linguistics, 2020. 4255–4264

153 Guo X, Zhao L, Qin Z, et al. Interpretable deep graph generation with node-edge co-disentanglement. In: Proceedings of Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, 2020. 1697–1707

154 Yang Y, Feng Z, Song M, et al. Factorizable graph convolutional networks. Advances in Neural Information Processing Systems, 2020, 33: 20286–20296

155 Li H, Wang X, Zhang Z, et al. Disentangled contrastive learning on graphs. Advances in Neural Information Processing Systems, 2021, 34: 21872–21884

156 Li H, Zhang Z, Wang X, et al. Disentangled graph contrastive learning with independence promotion. IEEE Transactions on Knowledge and Data Engineering, 2022

157 Li H, Wang X, Zhang Z, et al. Ood-gnn: Out-of-distribution generalized graph neural network. IEEE Transactions on Knowledge and Data Engineering, 2022

158 Karras T, Laine S, Aila T. A style-based generator architecturanetre for generative adversarial networks. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019

159 Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. In: Proceedings of International Conference on Learning Representations (ICLR), 2018

160 Sreekar P A, Tiwari U, Namboodiri A. Mutual information based method for unsupervised disentanglement of video representation. In: Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021. 6396–6403

161 Kim M, Lee H J, Lee S, et al. Robust video facial authentication with unsupervised mode disentanglement. In: Proceedings of 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020. 1321–1325

162 Ma J, Yu S. Human identity-preserved motion retargeting in video synthesis by feature disentanglement. arXiv preprint arXiv:2204.06862, 2022

163 Tang S, Mahjourian R, Zhao H, et al. Devias: Disentangled video representations of action and scene. In: Proceedings of ECCV, 2024

164 Colombo P, Piantanida P, Clavel C. A novel estimator of mutual information for learning to disentangle textual representations. In: Proceedings of Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021. 6539–6550

165 Hu Z, Yang Z, Liang X, et al. Toward controlled generation of text. In: Proceedings of International conference on machine learning. PMLR, 2017. 1587–1596

166 John V, Mou L, Bahuleyan H, et al. Disentangled representation learning for non-parallel text style transfer. In: Proceedings of Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 424–434

167 Zou Y, Liu H, Gui T, et al. Divide and conquer: Text semantic matching with disentangled keywords and intents. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2022, 2022. 3622–3632

168 Dougrez-Lewis J, Liakata M, Kochkina E, et al. Learning disentangled latent topics for twitter rumour veracity classification. In: Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021. 3902–3908

169 Zhu Q, Zhang W, Liu T, et al. Neural stylistic response generation with disentangled latent variables. In: Proceedings of Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021. 4391–4401

170 Zhang X, van de Meent J W, Wallace B C. Disentangling representations of text by masking transformers. In: Proceedings of Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021. 778–791

171 Zeng J, Jiang Y, Wu S, et al. Task-guided disentangled tuning for pretrained language models. In: Proceedings of Findings of the Association for Computational Linguistics: ACL 2022, 2022. 3126–3137

172 Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, 2019. 4171–4186

173 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning. PMLR, 2021. 8748–8763

174 Zhang Y, Wang X, Chen H, et al. Adaptive disentangled transformer for sequential recommendation. In: Proceedings of Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023. 3434–3445

175 Wang X, Pan Z, Zhou Y, et al. Curriculum co-disentangled representation learning across multiple environments for social recommendation. In: Proceedings of International conference on machine learning. PMLR, 2023

176 Zhang Z, Cui P, Zhu W. Deep learning on graphs: A survey. IEEE Transactions on Knowledge and Data Engineering, 2020

177 Zhou J, Cui G, Hu S, et al. Graph neural networks: A review of methods and applications. AI Open, 2020, 1: 57–81

178 Disentangled graph self-supervised learning for out-of-distribution generalization, author=Li, Haoyang and Wang, Xin and Zhang, Zeyang and Chen, Haibo and Zhang, Ziwei and Zhu, Wenwu, booktitle=International Conference on Machine Learning, pages=1–15, year=2024, organization=PMLR

179 Disentangled continual graph neural architecture search with invariant modular supernet, author=Zhang, Zeyang and Wang, Xin and Qin, Yijian and Chen, Hong and Zhang, Ziwei and Zhu, Wenwu, booktitle=International Conference on Machine Learning, pages=1–15, year=2024, organization=PMLR

180 Barrett D, Hill F, Santoro A, et al. Measuring abstract reasoning in neural networks. In: Proceedings of International conference on machine learning. PMLR, 2018. 511–520

181 Van Steenkiste S, Locatello F, Schmidhuber J, et al. Are disentangled representations helpful for abstract visual reasoning? Advances in neural information processing systems, 2019, 32

182 Locatello F, Poole B, Rätsch G, et al. Weakly-supervised disentanglement without compromises. In: Proceedings of International Conference on Machine Learning. PMLR, 2020. 6348–6359

183 Amizadeh S, Palangi H, Polozov A, et al. Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning". In: Proceedings of International Conference on Machine Learning. PMLR, 2020. 279–290

184 Xu J, Le H, Huang M, et al. Variational feature disentangling for fine-grained few-shot classification. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 8812–8821

185 Lin C C, Chu H L, Wang Y C F, et al. Joint feature disentanglement and hallucination for few-shot image classification. IEEE Transactions on Image Processing, 2021, 30: 9245–9258

186 Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: Proceedings of International conference on machine learning. PMLR, 2015. 1180–1189

187 Tokmakov P, Wang Y X, Hebert M. Learning compositional representations for few-shot recognition. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 6372–6381

188 Prabhudesai M, Lal S, Patil D, et al. Disentangling 3d prototypical networks for few-shot concept learning. arXiv preprint arXiv:2011.03367, 2020

189 Sun D, Ren T, Li C, et al. Learning to write stylized chinese characters by reading a handful of examples. arXiv preprint arXiv:1712.06424, 2017

190 Guo D, Tao D. Learning compositional representation for few-shot visual question answering. arXiv preprint arXiv:2102.10575, 2021

191 Dittadi A, Träuble F, Locatello F, et al. On the transfer of disentangled representations in realistic settings. In: Proceedings of International Conference on Learning Representations, 2020

192 Yoo H, Lee Y C, Shin K, et al. Disentangling degree-related biases and interest for out-of-distribution generalized directed network embedding. In: Proceedings of Proceedings of the ACM Web Conference 2023, 2023. 231–239

193 Zhang H, Zhang Y F, Liu W, et al. Towards principled disentanglement for domain generalization. In: Proceedings of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 8024–8034

194 Mu J, Qiu W, Kortylewski A, et al. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 13001–13011

195 Wu A, Liu R, Han Y, et al. Vector-decomposed disentanglement for domain-invariant object detection. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 9342–9351