解耦表征学习综述

文载道 [1, 2] 王佳蕊 [1, 2] 王小旭 [1, 2] 潘 泉 [1, 2]

摘 要 在大数据时代下, 以高效自主隐式特征提取能力闻名的深度学习引发了新一代人工智能的热潮, 然而其背后黑箱不可解释的 "捷径学习"现象成为制约其进一步发展的关键性瓶颈问题. 解耦表征学习通过探索大数据内部蕴含的物理机制和逻辑关系复杂性, 从数据生成的角度解耦数据内部多层次、多尺度的潜在生成因子, 促使深度网络模型学会像人类一样对数据进行自主智能感知, 逐渐成为新一代基于复杂性的可解释深度学习领域内重要研究方向, 具有重大的理论意义和应用价值. 本文系统地综述了解耦表征学习的研究进展, 对当前解耦表征学习中的关键技术及典型方法进行了分类阐述, 分析并汇总了现有各类算法的适用场景并对此进行了可视化实验性能展示, 最后指明了解耦表征学习今后的发展趋势以及未来值得研究的方向.

关键词 深度学习, 捷径学习, 潜在生成因子, 智能感知, 解耦表征学习

引用格式 文载道, 王佳蕊, 王小旭, 潘泉. 解耦表征学习综述. 自动化学报, 2022, **48**(2): 351−374

DOI 10.16383/j.aas.c210096

A Review of Disentangled Representation Learning

WEN Zai-Dao[1, 2] WANG Jia-Rui[1, 2] WANG Xiao-Xu[1, 2] PAN Quan[1, 2]

Abstract In the era of big data, deep learning has triggered the current rise of artificial intelligence which is known for its ability of efficient autonomous implicit feature extraction. However, the unexplainable "shortcut learning" phenomenon behind it has become a key bottleneck restricting its further development. By exploring the complexity of physical mechanism and logical relationship contained in big data, the disentangled representation learning aims to explore the multi-level and multi-scale explanatory generative latent factors behind the data, and prompts the deep neural network model to learn the ability of intelligent human perception. It has gradually become an important research direction in the field of deep learning, with huge theoretical significance and application value. This article systematically reviews the research of disentangled representation learning, classifies and elaborates state-of-the-art algorithms in disentangled representation learning, summarizes the applications of the existing algorithms and compares the performance of existing algorithms through experiments. Finally, the challenges and research trends in the field of disentangled representation learning are discussed.

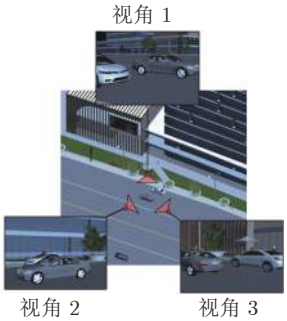
Key words Deep learning, shortcut learning, generative latent factors, intelligent perception, disentangled representation learning

Citation Wen Zai-Dao, Wang Jia-Rui, Wang Xiao-Xu, Pan Quan. A review of disentangled representation learning. *Acta Automatica Sinica*, 2022, **48**(2): 351−374

自动化系统, 大到复杂的导弹制导、自动驾驶、飞行控制等运动系统, 小到人脸图像识别、行人流量检测、视频跟踪监控等图像/视频解译系统, 均在国家、国防等重大生产、生活与管理进程中起到了不可替代的作用[1]. 随着人工智能技术最近几年的迅速发展, 采集数据的自动、精准智能感知对整个系统的智能辨识与控制预测能力至关重要, 备受研究者的广泛关注[2].

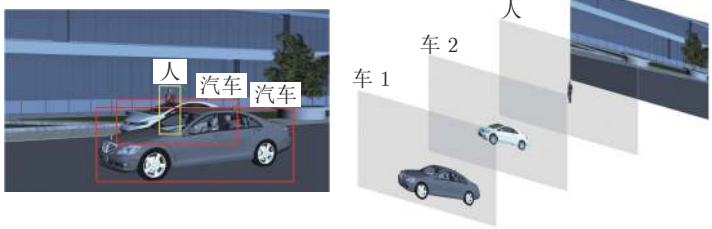
人类作为目前最为智能的生物系统, 能够通过各类生物传感器 (眼睛、鼻子、耳朵等) 接收周围环境的视觉、嗅觉、听觉等数据信号, 并将这些数据送入大脑进行融合处理, 挖掘出数据内部隐含的各类有效信息, 通过持续性学习将其汇总为简单的语义属性, 形成概念, 建立起抽象的逻辑关联规则, 最终结合自身具备的常识形成完整知识体系, 实现对各类复杂环境的智能化感知[3−4]. 例如, 将图 1 (a) 中从不同视角下拍摄得到的三幅不同交通图像作为视觉数据输入到人眼中, 人类便能够自主完成如下的层

收稿日期 2021-01-28 录用日期 2021-06-18
Manuscript received January 28, 2021; accepted June 18, 2021
国家自然科学基金 (61806165, 61790552, 61801020), 陕西省基础研究计划 (2020JQ-196) 资助
Supported by National Natural Science Foundation of China (61806165, 61790552, 61801020), the Natural Science Basic Research Plan in ShaanXi Province of China (2020JQ-196)
本文责任编委 王鼎
Recommended by Associate Editor WANG Ding
1. 西北工业大学自动化学院 西安 710129 2. 信息融合技术教育部重点实验室 西安 710129
1. School of Automation, Northwestern Polytechnical University, Xi′an 710129 2. Key Laboratory of Information Fusion Technology, Ministry of Education, Xi′an 710129



(a) 多个不同视角下的同一交通场景量测数据输入

(a) Measurement data input of the same traffic scene from multiple different views

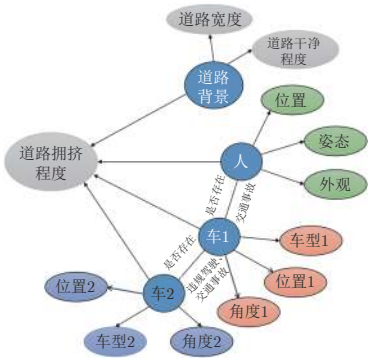


(b) 自动检测识别每幅视角图像中所含物体信息, 并想象还原出被遮挡物体的全貌

(b) Detecting and classifying the objects contained in every single image automatically, and visualizing the complete parts of the occluded objects

(c) 对于车辆而言, 将其内部属性解耦为车辆角度、车辆类型两类独立属性

(c) Disentangling the internal attributes of a vehicle into two independent parts: vehicle view and vehicle type



(d) 构建交通场景的智能树状知识体系图

(d) Constructing an intelligent knowledge graph of traffic scenes

(e) 人类想象泛化出发生车辆追尾碰撞现象的拥挤交通场景

(e) Humans can imagine a crowded traffic scene with a rear-end collision

图 1 人类对于交通场景量测数据的层次化智能感知示意图

Fig. 1 Humans′ hierarchical intelligent perception of a traffic scene

次化数据智能感知:

1) 检测并识别出图像中不同姿态、不同风格的物体, 并具有抗遮挡能力, 能够毫不费力地想象还原出被遮挡物体的全貌, 如图 1 (b) 所示;

2) 能够全面有效剖析出每类物体的各个内在属性并对该类物体进行全方位想象关联. 例如对于图 1 (c) 中的车辆而言, 假设将其内在属性认知为车型、角度两类, 人类便可按照这两类属性对已有图像进行相应的分组关联, 并能够通过组合不同的属性值想象出并未见过的车辆图像. 如此, 面对存在车辆的各类未知新场景, 人类能够不受大差异性视角或新型车辆的影响, 检测并识别出各类不同的车辆, 并能够精确推理出每辆车的内在属性值;

3) 能够结合一些常识推理 (例如两辆车相对位置过近或人躺在车辆行驶正中间的马路上时往往代表着交通事故的发生) 构建出代表不同对象间交互关系的树状知识体系图, 如图 1 (d) 所示. 利用该知识体系图, 人类能够通过对知识的改造重组想象泛化出各类符合因果逻辑关系的新场景, 例如图 1 (e) 中道路拥堵状态下的交通事故新场景. 该能力有助于人类对各类复杂场景进行因果知识关系梳理与认知更新, 从而轻松完成类似智能知识问答等复杂图像理解任务.

为了使现有系统真正实现对数据的自主智能感知, 借鉴人类这种层次化数据智能感知思想, 构建从数据、信息、语义、规则再到知识的多尺度、多层次、具有可解释性的数据表征至关重要.

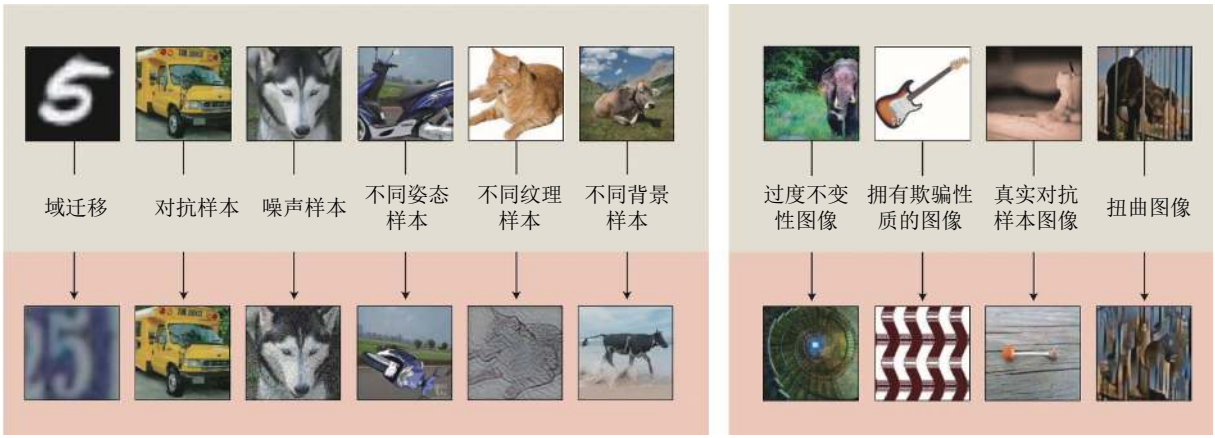
传统模式识别主要依据特定领域的专家经验知识进行显式的特征设计与推理, 从而完成相应任务. 随着误差反向传播 (Back propagation, BP) 人工神经网络的提出, 将传统专家知识驱动的显式特征提取方法替换为复杂数据驱动的神经网络隐式特征提取方法逐渐引起了学术界的关注. 尤其在 Hinton 等[5] 提出以深度神经网络为代表的深度学习技术

后, 相关以深度学习为主的隐式特征提取理论开始蓬勃发展, 逐渐在语音识别[6−8]、自然语言处理[9−11]、人脸识别[12−14]、目标检测[15−18] 等领域取得突破性进展. 截至目前, 深度学习技术已被广泛应用于多种复杂非线性系统的预测任务中[19]. 这类以提升特定预测任务性能指标为目的的判别式深度学习算法通过堆叠多层神经网络来构建从原始的输入数据到最终预测目标 (如物体类别、位置、姿态等) 的端到端非线性映射函数, 使机器能够从数据中自适应地进行学习, 有效缓解传统模式识别中手工设计选择显式特征的繁琐低效问题[20].

然而现有以有监督深度网络为代表的端到端黑箱判别式学习方法是一种捷径学习 (Shortcut learning) 策略[21−22], 即网络学习得到的判别性隐式抽象特征往往没有朝着人类所期望的方向进行泛化. 如图 2 所示, 对于图 2 (a) 中所显示的人类所具有的泛化能力并未被网络所学到. 与此相反, 在图 2 (b) 中, 网络学习得到的泛化能力又不能为人类所理解. 发生这种现象的本质原因在于现有判别式网络做出决策的评判标准仅仅为了提高训练样本数据的预测准确性. 在这种评判标准下, 网络会自主选择一条最容易、最精准地对训练集拟合的方向进行学习, 而这一方向并不一定是人类所期望网络学习的方向. 如图 3 所示, 网络学到得是所有决策空间中在训练集上展现出良好性能的一部分决策, 在这一部分决策内, 仅有一小部分决策能够泛化到服从独立同分布特性 (Independent and identically distributed, i.i.d) 的测试集上, 即图 3 中的蓝色区域. 然而人类真正期望网络做出的决策不仅能够泛化到 i.i.d 测试集上, 而且能够泛化到其余该分布以外 (Out-of-distribution, o.o.d) 的测试集中[23], 即图 3 中的红色区域部分. 现有大多数判别式网络仅旨在寻找蓝色区域内适应于 i.i.d 测试集的决策空间, 难以自主学到同时适应于 o.o.d 数据集的红色区域决策空间. 例如图 2 (a) 中, 当网络学习判断图像类别是否为猫时, 很容易聚焦于图像的纹理特征, 而忽略整体的形状特征, 这使得一幅具有猫的形状、大象纹理的图像会被网络判定为大象而不是猫; 又如图 2 (b)中, 网络对于一把吉他类别的判断可能仅在于评判其是否具有弯曲的纹理与线段等, 这使得该网络很容易将人类认为明显不是吉他的图像判定为吉他. 因此现有深度网络经常因为稳定性差、可解释性弱、易受欺骗攻击等饱受诟病[24−27].

为了缓解上述问题, 对网络学习方向施加一定的归纳偏好约束, 促使网络挖掘数据中所蕴含的常识推理与因果逻辑关系[28−31], 将有助于网络像人类一样学习从数据到信息到语义到规则再到知识的多尺度、多层次化数据表征. 基于此, 结合认知科学原理和视觉信息处理机制的解耦表征学习逐渐成为深度学习领域重要的研究方向[32−36]. 解耦表征学习旨在按照人类能够理解的方式从真实数据中对具有明确物理含义的生成因子 (如类别、位置、外观、纹理等) 进行解耦, 并给出其所对应的独立潜在表示, 引起国内外大量学者的广泛关注.

鉴于解耦表征学习深刻的理论意义, 所蕴含的应用价值以及可观的发展潜力, 本文对解耦表征学



(a) 对于人类而言相同类图像, 但对深度网络而言属于不同类图像示例图
(a) Image pairs assigned to the same category by humans, but not by DNNs

(b) 对于人类而言完全不相同类的图像, 但对深度网络而言属于相同类图像示例图
(b) Image pairs assigned to the same category by DNNs, but not by humans

图 2　深度网络的捷径学习 (Shortcut learning) 现象示例图[21]

Fig. 2　Examples of "Shortcut Learning" in DNNs[21]

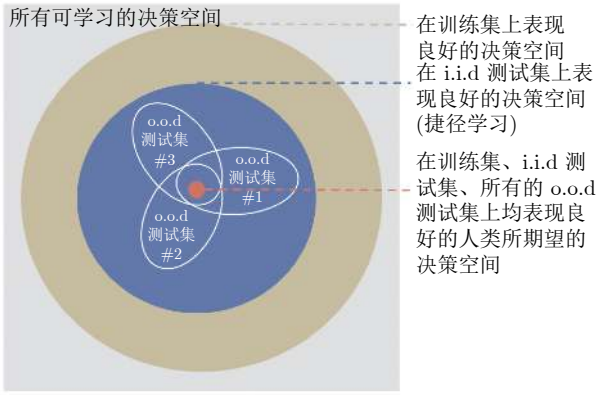


图 3 决策空间示意图[21]

Fig. 3 Taxonomy of decision rules[21]

习的研究进展进行了系统性的综述, 为进一步深入研究解耦表征学习机制、开发解耦表征学习应用潜力确立了良好的基础. 文中第 1 节对解耦表征学习基本概念、发展历史等进行了概述; 第 2 节着重介绍了从非结构化表征先验正则角度分析解耦表征学习最初的几种典型解决思路; 第 3 节则从结构化模型先验归纳偏好的角度挖掘模型架构设计对于现有解耦表征学习的启发; 第 4 节结合实际数据中所蕴含的物理知识对现有解耦表征学习研究进行进一步深入探索; 第 5 节则对前三节的模型算法进行对比分析论证. 最后, 在第 6 节指出了解耦表征学习未来的可能发展方向并对全文进行总结.

1 解耦表征学习

在表征学习中, 通常将真实数据 $\boldsymbol{x}$ 的生成过程建模为两部分: 从先验分布 $p(\boldsymbol{z})$ 中采样得到潜在变量取值 $\boldsymbol{z}$; 从条件数据生成分布 $p(\boldsymbol{x}|\boldsymbol{z})$ 中采样得到数据观测值 $\boldsymbol{x}$ [37]. 该模型背后的关键性假设在于将真实数据 $\boldsymbol{x}$ 视作由一系列物理语义可解释的因素 $\{v_1, v_2, \cdots, v_n\}$ 通过复杂未知的非线性系统映射函数 $Sim(\cdot)$ 作用相互耦合产生[37−38], 即 $\boldsymbol{x} = Sim(v_1, v_2, \cdots, v_n)$. 例如从宏观上看, 图 1 (b) 中的交通场景图像数据可看作由车 1、车 2、人、道路背景四个可解释的对象通过交通成像系统耦合而成, 图 1 (c) 中的车辆数据可看作由车型、角度两个可解释的生成因子通过车辆成像系统耦合而成. 从微观上看, 物质均由分子、原子等微观粒子耦合而成. 表征学习模型中的潜在变量 $\boldsymbol{z}$ 即为对这些物理可解释因子 $\{v_1, v_2, \cdots, v_n\}$ 的近似表征, 条件似然分布 $p(\boldsymbol{x}|\boldsymbol{z})$ 即为从概率角度对未知非线性系统映射函数 $Sim(\cdot)$ 的近似. 在此基础上, 解耦表征学习旨在学习可分离的潜在变量表示 $\boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n\}$, 且 $p(\boldsymbol{z}) = \prod_{k=1}^{n} p(\boldsymbol{z}_k)$, 使得该表示下每个潜在变量子集 $\boldsymbol{z}_k$ 能够对数据内部相对应的生成因子 v_k 进行有效表征控制.

传统解耦表征类研究可以追溯到独立成分分析 (Independent component algorithm, ICA) 方法[39−40], 旨在表示量测信号是如何由多种独立成分线性叠加而成. 据此构建的数学模型为:

$$\boldsymbol{x} = w_1\boldsymbol{z}_1 + w_2\boldsymbol{z}_2 + \cdots + w_n\boldsymbol{z}_n = \boldsymbol{W}\boldsymbol{z} \tag{1}$$

其中 $\boldsymbol{x}$ 为真实量测数据; $\boldsymbol{z}$ 为服从统计独立特性且非高斯分布的独立成分表示, 可视作潜在变量, 用于捕获影响线性系统输出的生成因子; $\boldsymbol{W}$ 为混合转换矩阵, 可近似为线性系统中将多输入生成因子线性叠加转换为量测输出数据的系统函数.

然而 ICA 一般仅适用于线性系统量测数据的解耦表征中, 对于复杂非线性系统量测数据的解耦表征类研究可通过将式 (1) 中用于近似线性系统函数的转换矩阵 $\boldsymbol{W}$ 替换为由多层参数化深度神经网络定义的复杂非线性转换函数, 如此便引申出自编码 (Auto-encoders, AE) 模型. 在自编码模型中, 由神经网络构成的编码器 h 对输入数据 $\boldsymbol{x}$ 进行编码形成潜在编码量 $\boldsymbol{z}$, 即 $\boldsymbol{z} = h(\boldsymbol{x})$; 另一个神经网络构成的解码器 f 则负责将这些潜在编码量 $\boldsymbol{z}$ 解码, 重构出原始数据 $\boldsymbol{x}$, 即 $\boldsymbol{x} = f(\boldsymbol{z}) = f(h(\boldsymbol{x}))$. 通过最小化重构误差, 自编码模型能够逐渐挖掘到对重构数据更有效的相关特征, 舍弃无关特征[41]. 该表征模型被 Schmidhuber[42] 于 1992 年用于非线性数据的解耦表征中, 他们建立的自适应预测器通过最小化可预测性原理惩罚每个潜在编码量所包含信息被其余潜在编码量预测出的概率来完成自编码模型中潜在编码量 $\boldsymbol{z}$ 的解耦任务.

现有大多数表征学习网络都是基于 Kingma 等提出的变分自编码 (Variational auto-encoders, VAE) 模型[43], 该模型从极大似然的角度对真实数据进行表征建模. 其中针对潜在变量 $\boldsymbol{z}$ 的推断过程, VAE 采用将真实数据输入到深度编码网络 f_ϕ 的方式进行变分近似后验推断 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$; 针对真实数据 $\boldsymbol{x}$ 的生成过程, VAE 采用将变分推断得到的潜在变量 $\boldsymbol{z}$ 输入到深度解码网络 g_θ 来近似数据生成建模 $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, 其中 $\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$. 对于整体 VAE模型中的网络参数 θ, ϕ 的求解优化方式采用极大对数似然思想, 如式 (2) 所示. 式中第一项为变分后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 与真实后验分布 $p(\boldsymbol{z}|\boldsymbol{x})$ 间的 KL 散度 (Kullback-Leibler divergence). 由于此项非负, 第二项 $\mathcal{L}(\theta, \phi; \boldsymbol{x}, \boldsymbol{z})$ 被称为真实数据 $\boldsymbol{x}$ 的变分下界, 代替式 (2) 成为 VAE 模型中新的优化目标函数, 如式 (3) 所示. 式中第一

项 $\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})$ 称为数据的条件对数似然项, 反映的是潜在变量 $\boldsymbol{z}$ 对于真实数据 $\boldsymbol{x}$ 的表征能力, 第二项 $\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z}))$ 常称为 KL项, 反映的是变分后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 与先验分布 $p(\boldsymbol{z})$ 间的相似性. 在 VAE 模型中, 由于人为选择的先验分布 $p(\boldsymbol{z})$ 通常满足独立特性, 如高斯正态分布等, 因此式 (3) 中的第二项 KL 项相当于对网络施加了一定程度的独立性约束, 通过该优化函数训练出的模型具备一定的解耦性能, 但实际应用过程中发现该约束能力还远不能实现对数据的有效解耦. 基于此问题, 目前大量学者通过在原始 VAE 中增添各类隐式或显式的归纳偏好促使网络学会数据内部各个可解释生成因子的有效解耦表征.

$$\max_{\theta,\phi} \ln p_\theta(\boldsymbol{x}) = \max_{\theta,\phi} \{\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_\theta(\boldsymbol{z}|\boldsymbol{x})) + \mathcal{L}(\theta,\phi;\boldsymbol{x},\boldsymbol{z})\} \tag{2}$$

$$\mathcal{L}(\theta,\phi;\boldsymbol{x},\boldsymbol{z}) = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \ln \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \ln p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})) \tag{3}$$

与 VAE 从贝叶斯的角度对真实数据进行生成分布建模不同, Goodfellow 等[44] 于 2014 年提出生成对抗网络 (Generative adversarial nets, GAN) 模型, 运用对抗学习思想在无需假设数据全概率生成分布模型的情况下正向模拟真实数据的生成过程. 具体而言, 该模型首先从人为假设的潜在先验分布 $p(\boldsymbol{z})$ 中采样, 近似复杂系统内影响数据输出的各个隐含生成因子; 随后将这些采样值送入用于模拟未知复杂系统函数的生成器 G 中, 输出生成的数据 $G(\boldsymbol{z})$; 最后采用判别器 D 对生成数据的真实性进行打分. 与 VAE 不同, GAN 不直接以数据分布与模型分布的差异作为目标函数, 而是采用对抗的方式, 先通过判别器去学习生成数据与真实数据的差异, 再引导生成器去缩小这种差异, 如式 (4) 所示, 逐渐寻找这种类似零和博弈中的纳什均衡解[45]. 相较于 VAE, GAN 不用对数据的分布模型进行显式设计, 避免了人为设计的复杂繁琐且赋予了网络更强大的生成数据能力. 然而 GAN 缺乏有效的推理机制, 只着重于数据的生成过程估计, 更适用于潜在因子已知情况下系统数据的近似生成问题, 而不是潜在因子的变化规律探索问题, 因此 GAN 模型难以直接应用到挖掘数据内部未知潜在生成因子的解耦表征研究中. 针对此, 目前大量学者提出 GAN 与 VAE 相结合的思想进一步开展对于真实数据的解耦表征学习研究.

$$\min_G \max_D V(D,G) = \mathrm{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\ln D(\boldsymbol{x})] + \mathrm{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\ln(1-D(G(\boldsymbol{z})))] \tag{4}$$

这两类生成式模型为解耦表征学习研究提供了许多新的思路. 然而在模型与数据都不存在归纳偏好 (Inductive bias) 的情况下, 网络无法自主无监督地学习出良好的解耦表征[37]. 对此, 大量学者针对表征变量、模型架构等提出了不同的归纳偏好设计, 促使模型拥有良好的解耦表征能力. 本文接下来将分别从非结构化表征先验归纳偏好、结构化模型先验归纳偏好、物理知识归纳偏好三方面对现有解耦表征学习的研究进展进行综述.

2 基于非结构化表征先验的解耦表征学习

在对真实数据进行解耦表征学习过程中, 对于潜在变量 $\boldsymbol{z}$ 的归纳偏好设计形式至关重要. 2016 年 ~ 2019 年, 大量解耦表征学习研究通过在原有生成式模型目标函数的基础上增添各类无监督正则项归纳偏好来为潜在变量 $\boldsymbol{z}$ 施加额外的独立性元先验约束, 促使网络偏向于学习满足独立统计分布特性的潜在变量表征. 本节将主要从独立性先验正则归纳偏好的角度出发, 对现有基于非结构化表征先验归纳偏好的解耦表征学习研究进行归纳整理分析.

对于 VAE 而言, 式 (3) 中的第二项 KL 项通过设计满足独立特性的先验分布 $p(\boldsymbol{z})$ 能够对网络学习到的变分后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 施加一定程度的独立性约束. 基于此, Higgins 等[46] 于 2017 年提出 β-VAE 模型, 直接对式 (3) 中的 KL 项施加大于一的罚项系数 β, 进而加强对近似后验分布的独立性约束, 鼓励网络着重学习潜在变量 $\boldsymbol{z}$ 的可分离性. 此时构成的新的优化函数如式 (5) 所示.

$$\mathcal{L}(\theta,\phi,\beta;\boldsymbol{x},\boldsymbol{z}) = \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \ln p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \beta\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z})) \tag{5}$$

其中 $\beta \geq 1$.

然而, 来自高 β 值的额外压力往往会使潜在变量所含的有效信息在经过解码器的过程中由于受约束的潜在瓶颈导致高频细节丢失, 对数据的表征能力下降, 难以达到数据的有效表征与解耦表征之间的最佳权衡. 基于此, 后续多项研究提出进一步的改进策略, 期望能够在不丢失过多数据表征能力的同时尽量提升潜在变量的解耦性能. Burgess 等[47]

于 2018 年从信息瓶颈理论分析的角度认为式 (5) 的对于近似后验分布的约束项为第一项重构项的信息瓶颈, 提出在训练过程中采用渐进策略逐渐增加潜在变量的信息容量, 如式 (6) 所示, 将有助于达成强表征能力与强解耦能力之间更好的权衡, 给予潜在变量更大的表示空间.

$$\mathcal{L}(\theta,\phi,C;\boldsymbol{x},\boldsymbol{z})=\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})-\gamma\left|\mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,||\,p(\boldsymbol{z})\right)-C\right| \tag{6}$$

其中 C 为随着网络训练不断线性增大的超参数.

除了上述将式 (3) 中的第二项 KL 正则项看作一个整体进行改动以提高网络解耦表征的能力外, Makhzani 等[48] 提出对抗自编码 (Adversarial auto-encoders, AAE), 对式 (3) 中的 KL 项进行更加细致化与规范化的推导, 如式 (7) 所示. 他们认为式 (7) 中第三项互信息项反映的是潜在变量 $\boldsymbol{z}$ 与输入数据 $\boldsymbol{x}$ 间的相关性. 若惩罚该项, 将有可能导致潜在变量 $\boldsymbol{z}$ 无法对输入数据 $\boldsymbol{x}$ 进行有效表征. 而式中第二项有关累积后验分布与先验分布的 KL 项则是真正有助于提升解耦效能的关键项. 基于此, 他们采用对抗约束的方式仅惩罚式 (7) 中第一项重构项与第二项 KL 项. 该分解相较于 β-VAE将式 (7) 中后两项看作一个整体进行惩罚, 更好地达到数据解耦性能与表征性能间的平衡. Kumar等[49] 认为 AAE 在运用对抗思想的同时会面临对抗训练所存在的鞍点等问题[50]. 他们提出的 DIP-VAE (Disentangled inferred prior variational auto-encoders) 模型将潜在变量后验累积分布 $q_\phi(\boldsymbol{z})$ 与先验分布 $p(\boldsymbol{z})$ 均假设为高斯分布, 利用矩估计思想设计了两种矩匹配项来对后验分布的协方差矩阵进行约束来促使二者分布达到一致, 其设计形式如式 (8)、(9) 所示. 该方法相对于 AAE 而言大大简化了训练过程, 避免了对抗训练中所可能出现的鞍点等问题.

$$\begin{aligned}&\mathcal{L}(\theta,\phi;\boldsymbol{x},\boldsymbol{z})=\\&\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})-\mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,||\,p(\boldsymbol{z})\right)=\\&\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})-\\&\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\ln\left(\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}\right)\right]=\\&\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})-\\&\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\ln\frac{q(\boldsymbol{z})}{p(\boldsymbol{z})}+\ln\frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z})}\right]=\\&\underbrace{\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})}_{\text{reconstruction term}}-\underbrace{\mathrm{KL}\left(q_\phi(\boldsymbol{z})\,||\,p(\boldsymbol{z})\right)}_{\text{marginal KL}}-\underbrace{I(\boldsymbol{z};\boldsymbol{x})}_{\text{mutual info}}\end{aligned} \tag{7}$$

$$\max_{\theta,\phi}ELBO(\theta,\phi)-\lambda_{od}\sum\nolimits_{i\neq j}\left[Cov_{p(\boldsymbol{x})}\left[\mu_\phi(\boldsymbol{x})\right]\right]_{ij}^2-\lambda_d\sum\nolimits_i\left(\left[Cov_{p(\boldsymbol{x})}\left[\mu_\phi(\boldsymbol{x})\right]\right]_{ii}-1\right)^2 \tag{8}$$

$$\max_{\theta,\phi}ELBO(\theta,\phi)-\lambda_{od}\sum\nolimits_{i\neq j}\left[Cov_{q_\phi(\boldsymbol{z})}\left[\boldsymbol{z}\right]\right]_{ij}^2-\lambda_d\sum\nolimits_i\left(\left[Cov_{q_\phi(\boldsymbol{z})}\left[\boldsymbol{z}\right]\right]_{ii}-1\right)^2 \tag{9}$$

然而当潜在变量先验分布 $p(\boldsymbol{z})$ 设计有偏差时, 采用上述方法对后验累积分布 $q_\phi(\boldsymbol{z})$ 与先验分布 $p(\boldsymbol{z})$ 施加强一致性约束会导致数据表征学习的有效性减弱. 基于此, Kim 等[51] 与 Chen 等[52] 先后于 2018 年提出能够直接鼓励后验累积分布 $q(\boldsymbol{z})$ 服从因式阶乘分布的惩罚项: $\mathrm{KL}(q(\boldsymbol{z})||\prod_{i=1}^d q(z_i))$项. 其中 Kim 等[51] 所提出的 Factor-VAE 直接在原始 VAE 优化函数中增加该惩罚项, 如式 (10) 所示, 用于提升模型的解耦性能. Chen 等[52] 所提出的 β-TCVAE (Total correlation variational auto-encoders) 从理论推导角度将式 (7) 中第二项 $\mathrm{KL}(q(\boldsymbol{z})||p(\boldsymbol{z}))$ 进一步分解, 如式 (11) 所示. 进而通过对不同项赋予不同的权重值构成新的优化函数, 如式 (12) 所示. 两种方法对于 $\mathrm{KL}(q(\boldsymbol{z})||\prod_{i=1}^d q(z_i))$ 项的相似性度量均采用对抗方式求解.

$$\begin{aligned}&\mathcal{L}(\theta,\phi,\gamma;\boldsymbol{x},\boldsymbol{z})=\\&\quad\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})-\mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,||\,p(\boldsymbol{z})\right)-\\&\quad\gamma\mathrm{KL}(q(\boldsymbol{z})||\prod\nolimits_{i=1}^d q(z_i))\end{aligned} \tag{10}$$

$$\mathrm{KL}\left(q(\boldsymbol{z})\,||\,p(\boldsymbol{z})\right)=\mathrm{KL}(q(\boldsymbol{z})||\prod\nolimits_{i=1}^d q(z_i))+\sum_j\mathrm{KL}(q(z_j)||p(z_j)) \tag{11}$$

$$\begin{aligned}&\mathcal{L}(\theta,\phi,\alpha,\beta,\gamma;\boldsymbol{x},\boldsymbol{z})=\\&\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})-\beta\mathrm{KL}\left(q(\boldsymbol{z})\,||\prod\nolimits_{i=1}^d q\left(z_i\right)\right)-\\&\alpha I(\boldsymbol{z};\boldsymbol{x})-\gamma\sum_j\mathrm{KL}(q(z_j)||p(z_i))\end{aligned} \tag{12}$$

除了上述基于表征独立分布特性设计先验正则用于解耦表征学习外, 还有部分学者从其余表征分布特性的角度出发对上述方法进行了进一步的补充. 以下将分别从离散型潜在变量分布特性、与数据相关的潜在变量解耦特性、序列图像中潜在变量的时空一致性以及潜在变量的稀疏性四个角度进行展开描述.

用于捕捉数据内部生成因子的潜在变量除了类似位置、外观等连续型潜在变量外, 还存在着类别

等离散型潜在变量. 这类离散型潜在变量的存在会使得深度网络在进行梯度回传时出现无法有效求微的难解问题. 基于此, Dupont 等[53] 提出 JointVAE, 使用连续的 Concrete 分布[54] 来对离散型潜在变量进行建模, 并采用连续型潜在变量 $\boldsymbol{z}$ 与离散型潜在变量 $\boldsymbol{c}$ 联合分布建模 $q_\phi(\boldsymbol{z},\boldsymbol{c}|\boldsymbol{x})$ 的方式, 将式 (6) 中的目标函数扩展为式 (13) 形式, 为涉及到离散型潜在变量的解耦问题提供了一个很好的思路.

$$\begin{aligned}\mathcal{L}(\theta,\phi)=&\,\mathrm{E}_{q_\phi(\boldsymbol{z},\boldsymbol{c}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{c})-\\&\gamma\left|\mathrm{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,||\,p(\boldsymbol{z})\right)-C_z\right|-\\&\gamma\left|\mathrm{KL}\left(q_\phi(\boldsymbol{c}|\boldsymbol{x})\,||\,p(\boldsymbol{c})\right)-C_c\right|\end{aligned}\tag{13}$$

此外, 在潜在变量子集中还会存在着一些无关噪声干扰. 对此, Chen 等[55] 于 2016 年提出生成对抗网络的信息论扩展网络 InfoGAN (Information maximizing generative adversarial nets). 该网络旨在将潜在变量解耦为不可压缩噪声源 $\boldsymbol{z}$ 与有效信息源 $\boldsymbol{c}$ 两部分. 考虑到有效信息源 $\boldsymbol{c}$ 应该在数据生成过程中发挥主要作用, InfoGAN 提出最大化有效信息源 $\boldsymbol{c}$ 与生成数据 $G(\boldsymbol{z},\boldsymbol{c})$ 间的互信息 $I(\boldsymbol{c};G(\boldsymbol{z},\boldsymbol{c}))$. 该文献使用变分后后验分布 $Q(\boldsymbol{c}|\boldsymbol{x})$ 来近似真实后验分布 $P(\boldsymbol{c}|\boldsymbol{x})$, 设计出一种可以有效优化的互信息目标下界 $\mathcal{L}_I(G,Q)$, 将难解问题可解化. 其定义如式 (14) 所示. 将其并入 GAN 的优化目标函数中, 如式 (15) 所示, 旨在鼓励网络学习更具可解释性和有意义的表征形式.

$$\begin{aligned}&\mathcal{L}_I(G,Q)=\\&\mathrm{E}_{\boldsymbol{c}\sim P(\boldsymbol{c}),\boldsymbol{x}\sim G(\boldsymbol{z},\boldsymbol{c})}\left[\ln Q(\boldsymbol{c}|\boldsymbol{x})\right]+H(\boldsymbol{c})=\\&\mathrm{E}_{\boldsymbol{x}\sim G(\boldsymbol{z},\boldsymbol{c})}\left[\mathrm{E}_{\boldsymbol{c}'\sim P(\boldsymbol{c}|\boldsymbol{x})}\left[\ln Q(\boldsymbol{c}'|\boldsymbol{x})\right]\right]+H(\boldsymbol{c})\le\\&I(\boldsymbol{c};\,G(\boldsymbol{z},\boldsymbol{c}))\end{aligned}\tag{14}$$

$$\begin{aligned}&\min_{G,Q}\max_{D}V_{\mathrm{InfoGAN}}(D,G,Q)=\\&\quad V(D\,,G)-\lambda\mathcal{L}_I(G,Q)\end{aligned}\tag{15}$$

除此之外, Kim 等[56] 认为对于这些无关噪声干扰间的解耦程度并不需要额外约束, 于 2019 年引入相关性指标 $\boldsymbol{r}$ 对 Factor-VAE 进行改进, 提出 RF-VAE (Relevance factor variational auto-encoders). 如式 (16) 所示, 旨在使式 (10) 中的最后一项仅作用于对数据有用的相关潜在变量.

$$\begin{aligned}\mathcal{L}(\theta,\phi)=&\,\mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\ln p_\theta(\boldsymbol{x}|\boldsymbol{z})+\\&\mathrm{E}_{p_d(\boldsymbol{x})}\left[\sum_{i=1}^{d}\lambda_j\mathrm{KL}(q(z_j|\boldsymbol{x})||p(z_j))\right]+\\&\gamma\mathrm{KL}(q(\boldsymbol{r}\circ\boldsymbol{z})||\textstyle\prod_i q(r_i\circ z_i))+\eta\|\boldsymbol{r}\|_1\end{aligned}\tag{16}$$

对于序列型数据而言, 时空一致性是其本征重要物理特性之一, 若在潜在变量分布建模时融入该特性将有助于网络学习到更符合真实物理规律的表征形式. 基于此, Grathwohl 等[57] 于 2016 年针对视频序列中相对静止的背景场景与随时间平缓运动的前景目标间的解耦表征问题, 提出采用式 (17) 的形式对背景表征分布进行建模, 采用式 (18) 的形式对前景目标表征分布进行建模. 此种建模方式模拟了真实世界中背景时空不变性与前景运动目标时空平缓变化的特征, 更为有效合理.

$$\begin{aligned}p(h_0)&=\mathrm{N}(0,\boldsymbol{I})\\p(h_t)&=\mathrm{N}(h_0,\sigma_t\cdot\boldsymbol{I})\end{aligned}\tag{17}$$

$$\begin{aligned}p(h_0)&=\mathrm{N}(0,\boldsymbol{I})\\p(h_t)&=\mathrm{N}(h_{t-1},\sigma_t\cdot\boldsymbol{I})\end{aligned}\tag{18}$$

除了上述提及的几个特性外, 还应注意到数据内部潜在变量表征往往是具有稀疏特性的, 即不是每个潜在变量都需要对数据进行表征. 传统变分自编码模型对于潜在变量的先验建模大多采用高斯正态分布, 难以反映其内有稀疏特性, 而学生 t 分布、拉普拉斯分布等厚尾分布则可以很好地体现变量的稀疏分布特性. 基于此, Kim 等[58] 于 2019 年提出分层贝叶斯深度变分自编码模型, BF-VAE (Bayes factor variational auto-encoders). 同 InfoGAN[55] 一样, 将潜在变量分为相关潜在变量与干扰潜在变量两类. 他们认为厚尾分布更适用于相关潜在变量的分布建模, 而传统高斯分布则适用于干扰潜在变量的分布建模. 利用此思想, 在传统高斯先验的方差上引入超先验的同时保持传统 VAE 的易学性与推理性, 将 VAE 扩展为分层贝叶斯模型.

因此针对具体问题, 应具体分析其背后所具备的物理分布特性, 并基于此选择适用的表征分布模型, 将有助于提升整体网络的解耦表征学习能力. 此外, 对于本节所涉及的各类先验正则化归纳偏好方法的汇总如表 1 所示. 从表 1 中可看出, 本节所涉及的各类算法虽然一定程度上能够实现数据的有效解耦表征, 但这类算法的学习过程依旧缺乏明确的物理语义导向. 这将进一步引出本文后两节基于结构化模型先验归纳偏好与基于物理知识先验归纳偏好的解耦表征学习类研究算法探讨.

3 基于结构化模型先验归纳偏好的解耦表征学习

对于第 2 节中基于非结构化表征先验的解耦表征学习方法, Montero 等[59] 于 2021 年设计实验, 调整数据集中的部分属性取值范围, 分别测试了原始

表 1 非结构化表征先验归纳偏好方法对比

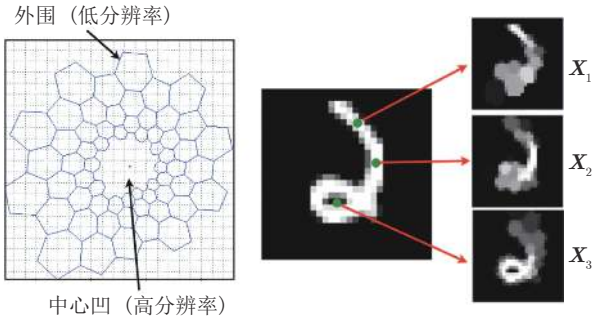
Table 1 Comparison of unstructured representation priori induction preference methods

工作	正则项	优点	缺点
β-VAE[46]	$-\beta D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}\|\boldsymbol{x}) \,\|\|\, p(\boldsymbol{z})\right)$	高 β 值促使网络所学到的后验分布与先验分布尽可能服从相似的独立统计特性, 提升解耦性能.	高 β 值在提升解耦性能的同时会限制网络的数据表征能力, 直观反映为重构性能降低, 无法很好权衡二者.
Understanding disentangling in β-VAE[47]	$-\gamma\left\|\mathrm{KL}\left(q(\boldsymbol{z}\|\boldsymbol{x}) \,\|\|\, p(\boldsymbol{z})\right) - C\right\|$	从信息瓶颈角度分析 β-VAE, 在训练过程中渐进增大潜在变量的信息容量 C, 能够在一定程度上改善了网络对于数据表征能力与解耦能力间的权衡.	该设计下的潜在变量依旧缺乏明确的物理语义, 且网络增加了信息容量 C 这一超参数, 需要人为设计其渐进增长趋势.
Joint-VAE[53]	$-\gamma\left\|\mathrm{KL}\left(q_\phi(\boldsymbol{z}\|\boldsymbol{x}) \,\|\|\, p(\boldsymbol{z})\right) - C_z\right\|$ $-\gamma\left\|\mathrm{KL}\left(q_\phi(\boldsymbol{c}\|\boldsymbol{x}) \,\|\|\, p(\boldsymbol{c})\right) - C_c\right\|$	运用 Concrete 分布[54] 解决离散型潜在变量的解耦问题.	潜在变量缺乏明确物理语义.
AAE[48]	$D_{\mathrm{JS}}\left[\mathrm{E}_\phi\left(\boldsymbol{z}\right)\|\|p\left(\boldsymbol{z}\right)\right]$	利用对抗网络完成累积后验分布与先验分布间的相似性度量, 使得潜在变量的表达空间更大, 表达能力更强.	面临对抗网络所存在的鞍点等训练问题[50].
DIP-VAE[49]	$-\lambda_{od}\sum_{i\neq j}\left[Cov_{q_\phi(\boldsymbol{z})}\left[\boldsymbol{z}\right]\right]_{ij}^2$ $-\lambda_d\sum_i\left(\left[Cov_{q_\phi(\boldsymbol{z})}\left[\boldsymbol{z}\right]\right]_{ii}-1\right)^2$	设计更简便的矩估计项替代 AAE[48] 中对抗网络的设计, 计算更为简洁有效.	该设计仅适用于潜在变量服从高斯分布的情况且并未限制均值矩或更高阶矩, 适用范围有限.
Factor-VAE[51]	$D_{\mathrm{JS}}(q(\boldsymbol{z})\|\|\prod_{i=1}^d q(z_i))$	设计对抗网络直接鼓励累积后验分布 $q(\boldsymbol{z})$ 服从因子分布, 进一步改善了网络在强表征能力与强解耦能力间的权衡.	面临对抗网络所存在的鞍点等训练问题[50].
RF-VAE[56]	$D_{\mathrm{JS}}(q(\boldsymbol{r}\circ\boldsymbol{z})\|\|\prod_{i=1}^d q(r_i\circ z_i))$	引入相关性指标 $\boldsymbol{r}$ 使得网络对于无关隐变量间的解耦程度不作约束.	相关性指标 $\boldsymbol{r}$ 也需要由网络学习得到, 加深了网络训练的复杂性.
β-TCVAE[52]	$-\alpha I_q(\boldsymbol{x};\boldsymbol{z})-$ $\beta\mathrm{KL}\left(q\left(\boldsymbol{z}\right)\|\|\prod_{i=1}^d q\left(z_i\right)\right)$ $-\gamma\sum_j \mathrm{KL}(q(z_j)\|\|p(z_i))$	证明了TC总相关项 $\mathrm{KL}(q(\boldsymbol{z})\|\|\prod_{i=1}^d q(z_i))$ 的重要性并赋予各个正则项不同的权重值构成新的优化函数使其具有更强的表示能力.	引入更多的超参需要人为调试.

VAE, β-VAE, Factor-VAE 三类方法在相应测试集上的泛化性能, 发现这类单纯施加非结构化表征先验正则归纳偏好的方法对于模型学习方向的约束能力隐形且较弱, 不足以支持复杂情境设计下的组合泛化性. 他们认为设计模块化、结构化、融合实际物理机理的模型尤为重要. 本节将从顺序深度递归网络、层次深度梯形网络以及树形网络三个由人类认知过程所启发的高度显式结构化网络模型来对现有基于结构化模型先验归纳偏好的解耦表征学习类研究进行归纳探讨, 对于融入实际物理机理的模型设计将在本文第 4 节中进行探讨.

3.1 顺序深度递归网络

目前, 大多数基于深度学习进行图像理解的方法往往倾向于一次性理解整幅场景. 在生成式神经网络的背景下, 这通常意味着所有像素都受单次潜在分布的约束, 且网络无法进行迭代自校正. 然而人类进行场景感知时往往不倾向于同时处理整幅场景. 相反, 人类会利用连续的中心凹运动进行 "主动感知": 在给定的时间内, 有选择地将注意力集中在中心凹的高分辨率视觉空间中, 并随着时间的推移将来自不同注视点的信息结合起来, 指导未来的眼球中心凹运动序列 (旋转和平移) 决策, 逐渐建立起整幅场景的全面表征, 如图 4 所示. 受到该人类



(a) 视网膜瞥视示意图
(a) Illustration of the retina glance

(b) 视网膜瞥视过程可视化示例图
(b) Examples of glimpses computed by the retinal transformation

图 4 人类视网膜瞥视过程图[60]

Fig. 4 Illustration of the retinal transformation[60]

感知方法的强烈驱动, 许多学者逐渐发现 "一次性感知" 表示方法从根本上很难扩展到大图像或目标占比过小的图像场景. 与此相比, 通过一系列的局部瞥视或显著区域捕捉可以更好地捕获视觉结构[60−62], 这种思想可以通过使用递归神经网络执行概率迭代推理来实现, 使得网络每次只关注部分图像进行处理, 最终整合至整幅图像. 这种顺序递归模型的明显优点是, 通过将复杂数据分布映射到一系列更简单的问题中, 反复生成以先前状态为条件

的输出, 简化了建模复杂数据分布的问题. 然而该方法的难点在于如何选择注意机制以及如何将显著区域的位置与提取的特征相结合, 如何选择递归次数等.

Larochelle 等[60] 于 2010 年首先提出一种特殊的模拟人眼中心凹特性的受限玻尔兹曼机模型, 该模型在可见单元 (瞥视), 隐藏单元 (累积特征) 以及控制可见单元与隐藏单元连接的位置相关单元间建立三阶连接, 学习如何在多个固定点上累积有关单个目标形状的信息. 基于此思想, Mnih 等于 2014 年[61] 将注意机制问题看作是以目标为导向的智能体与视觉环境交互的顺序决策过程. 他们提出基于递归神经网络的循环注意机制模型, 为每次决策设计计算一个标量奖励的反馈, 从而结合强化学习的训练策略, 促使最终决策的总和最大化. 该模型随后被 Gregor 等[62] 扩展为深度递归视觉注意模型 (DRAW) 用于生成图像, 在 VAE 的框架下采用递归循环网络来构建编码器与解码器, 每次循环通过解码器发出的修改累积迭代地构造场景, 同时嵌入空间二维高斯滤波器来产生位置、缩放平滑变化的局部图像 “块”充当每次迭代过程中网络所选定的注意区域. 而对于如何选择迭代次数, 他们将其视为人为提前设定的固定超参数.

目前该思想被广泛用于解决复杂场景多目标解耦问题. 对于场景的认知, Henderson 等[63] 给出了以下定义: “场景是真实世界环境的语义连贯 (通常是可命名的) 视图, 包含背景元素和以空间特定方式排列的多个离散对象.” 基于此, 许多学者将循环递归网络每次的迭代过程视为新目标的形成过程, 并在每次形成新目标后通过一组特定的仿射函数将其与之前场景相复合. 其中, Eslami 等[64] 于 2016 年提出的基于 VAE 的结构化图像模型 AIR (Attend-infer-repeat) 引起人们的广泛关注, 后续被大量引用于复杂多目标场景的解耦表征研究中. 该模型可理解为基于对象的解耦表征, 通过将编码推理网络构建为递归神经网络的形式促使网络迭代学习关于场景中存在的每个对象的解耦表征. 且由于该模型将对象表示为 {存在概率、特有属性、坐标} 三类, 该模型可被用于目标检测、识别等下游推理任务中. 后续被 Crawford 等[65] 改进为适用于较多目标场景的检测模型 SPAIR (Spatially invariant attend-infer-repeat). AIR 的整体架构设计如图 5 所示, 通过平摊、迭代推理的方式来逐目标地实现多对象场景的理解, 并结合空间仿射变换对坐标这一潜在表征施加强物理约束, 有效指引了网络的学习方向. 除此之外, AIR 将网络的迭代次数, 即前景目标个数也视为一个隐变量, 服从特定的分布, 这一方法对于可变数量的前景目标检测具有更强的鲁棒性与泛化性. 然而该方法只能处理简单背景下少量前景目标的检测等问题, 且并未进一步考虑不同目标间的语义关联关系.

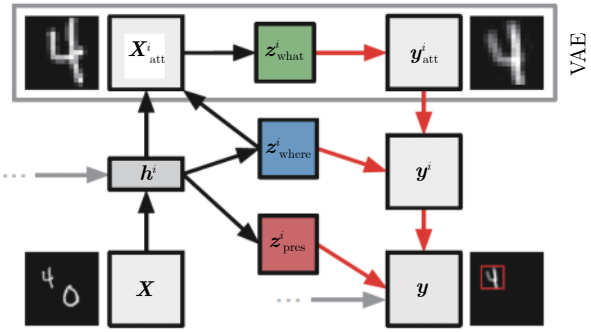


图 5 模型架构设计图[64]

Fig. 5 AIR framework[64]

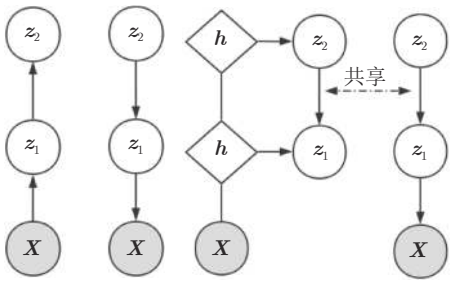
为了扩展 AIR 在连续视频场景下的使用, Kosiorek 等[66] 于 2018 年提出的 SQAIR (Sequential attend-infer-repeat) 将视频中所具有的时空一致性加入原始 AIR 模型中进行改进. 具体而言, 该模型将视频数据的生成过程分为两支路实现: 传播支路 (Propagation, PROP) 用于负责更新 (或遗忘) 前一时间步中所含对象的潜在变量观测 (图像), 且结合关系 RNN[67] 来对物体间的关系进行捕获; 发现支路 (Discovery, DISC) 在传播支路 (PROP) 的基础上进一步检测是否有新对象出现. 该模型能够实现简易视频数据集中的目标检测跟踪问题, 为具有时序变化性的变分自编码架构设计提供了前瞻性的解决思路. 此外, Massague 等[68] 于 2020 年提出视频部分帧缺失情况下的解耦表征问题. 他们认为人类在视频帧突然缺失或突然受到干扰的情况下自然而然地认为之前帧中的物体依旧存在, 且其运动轨迹遵循之前的规律. 基于此, 他们在潜在空间设计中多考虑了一组代表缺失状态的潜在变量子集用于判别当前帧数据的缺失状态, 若缺失, 则通过在过去帧的潜在空间采样来插补近似缺失帧的潜在表征. 此设计促使网络自监督地学习缺失数据的插补表征方式, 一定程度上解决了视频部分帧缺失数据的解耦表征问题.

3.2 层次深度梯形网络

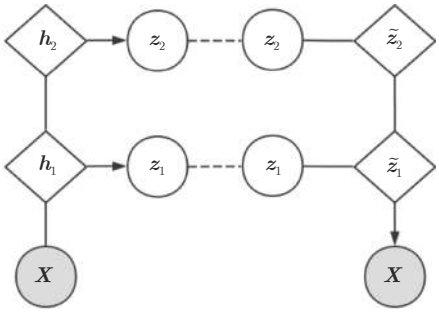
除了第 3.1 节中利用循环递归网络实现顺序迭代逐步处理特定任务外, 考虑到现实世界中许多自然信号本身所特有的成分分层特性, 本节集中于层次深度梯形网络的设计搭建, 赋予深度网络不同语

义特征提取过程中显式层次结构的归纳偏好, 即通过组合较低层的语义特征来获得较高层的语义特征表示. 例如在现实世界中, 边缘的局部组合形成图案, 图案组装形成零件, 零件组装形成对象.

Sønderby 等[69] 于 2016 年提出梯形变分自编码网络 (Ladder variational auto-encoders, LVAE). 与传统 VAE 所使用的推理模型与生成模型间无交互作用的纯自底向上推理过程 (如图 6 (a) 所示) 不同, 该文献提出推理与生成模型中共享自顶向下的依赖结构, 如图 6 (b) 所示, 使得模型的推理过程只用简单修正生成分布, 将优化过程变得更加容易.



(a) VAE[43] (b) LVAE[69]



(c) VLAE[70]

图 6 深度梯形网络模型图

Fig. 6 Deep ladder network models

然而 Zhao 等[70] 认为 LVAE[69] 在训练到最优的情况下, 仅底层潜在变量包含足够的信息用于重建数据分布, 其余层则很容易被忽略. 且通常用于构建层次生成模型的许多构建块不太可能帮助学习解耦特征. 基于此, 他们于 2017 年提出变分梯形自动编码网络 (Variational ladder auto-encoders, VLAE), 通过在每一层潜在变量与图像之间映射所需的计算程度来分离图像的潜在变量子集. 将不同层次的潜在变量与具有不同表达能力 (深度) 的网络连接起来; 鼓励模型在顶部放置高层次、抽象的特征 (如身份特征等), 在底部放置低层次、简单的特征 (如边缘特征等). 该模型设计如图 6 (c) 所示, 其中条件生成模型 $p(\boldsymbol{x}|\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_L)$ 被隐式定义为式 (19) 的形式. 推理网络定义为式 (20) 的形式. 这种模型设计使得越高层、越抽象的潜在表示需要越复杂的网络来捕获, 在不需要特定任务规则化或先验知识的情况下, 能够学习到高度可解释的、解耦的层次特征. 该模型随后被 Willetts 等[71] 用于促使网络在不同层解耦代表不同属性的潜在表征, 从而基于该表征在各个层次实现按照不同属性区分的聚类任务, 他们称其为解耦聚类. 除此之外, Esmaeili 等[72] 于 2019 年从多级隐变量角度出发, 将潜在变量分为组间潜在变量与组内潜在变量两级来对 KL 项进行进一步分解, 如式 (21)、(22) 所示, 提出了基于 VAE 的两级分层 HFVAE (Hierarchically factorized variational auto-encoders) 模型. 该模型可以通过控制两级隐变量不同的权重系数来控制组间隐变量与组内隐变量的相对解耦程度, 如式 (23) 所示.

$$\begin{aligned} \tilde{\boldsymbol{z}}_L &= f_L(\boldsymbol{z}_L) \\ \tilde{\boldsymbol{z}}_l &= f_l(\tilde{\boldsymbol{z}}_{l+1}, \boldsymbol{z}_l) \\ \boldsymbol{x} &\sim r(\boldsymbol{x}; f_0(\tilde{\boldsymbol{z}}_1)) \end{aligned} \tag{19}$$

$$\begin{aligned} \boldsymbol{h}_l &= g_l(\boldsymbol{h}_{l-1}) \\ \boldsymbol{z}_l &\sim \mathrm{N}(\mu_l(\boldsymbol{h}_l), \sigma_l(\boldsymbol{h}_l)) \end{aligned} \tag{20}$$

其中 f_l, g_l 均为非线性神经网络映射.

$$\begin{aligned} &\mathrm{KL}(q_\phi(\boldsymbol{z})||p(\boldsymbol{z})) = \\ &\mathrm{E}_{q_\phi(\boldsymbol{z})}\left[\ln\left[\frac{q_\phi(\boldsymbol{z})}{\prod_d q_\phi(\boldsymbol{z}_d)} \circ \frac{\prod_d q_\phi(\boldsymbol{z}_d)}{\prod_d p(\boldsymbol{z}_d)} \circ \frac{\prod_d p(\boldsymbol{z}_d)}{p(\boldsymbol{z})}\right]\right] = \\ &\mathrm{E}_{q_\phi(\boldsymbol{z})}\left[\ln\frac{q_\phi(\boldsymbol{z})}{\prod_d q_\phi(\boldsymbol{z}_d)} - \ln\frac{p(\boldsymbol{z})}{\prod_d p(\boldsymbol{z}_d)}\right] + \\ &\sum\nolimits_d \mathrm{KL}(q_\phi(\boldsymbol{z}_d)||p(\boldsymbol{z}_d)) \end{aligned} \tag{21}$$

其中

$$\begin{aligned} &\mathrm{KL}(q_\phi(\boldsymbol{z}_d)||p(\boldsymbol{z}_d)) = \\ &\quad \mathrm{E}_{q_\phi(\boldsymbol{z})}\left[\ln\frac{q_\phi(\boldsymbol{z}_d)}{\prod_e q_\phi(\boldsymbol{z}_{d,e})} - \ln\frac{p(\boldsymbol{z}_d)}{\prod_e p(\boldsymbol{z}_{d,e})}\right] + \\ &\quad \sum\nolimits_e \mathrm{KL}(q_\phi(\boldsymbol{z}_{d,e})||p(\boldsymbol{z}_{d,e})) \end{aligned} \tag{22}$$

$$\begin{aligned} &\mathcal{L}(\theta, \phi; \boldsymbol{x}, \boldsymbol{z}) = \\ &\quad \mathrm{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \ln p(\boldsymbol{x}|\boldsymbol{z}) - \alpha I_{q_\phi}(\boldsymbol{x}; \boldsymbol{z}) - \\ &\quad \beta \mathrm{E}_{q_\phi(\boldsymbol{z})}\left[\ln\frac{q_\phi(\boldsymbol{z})}{\prod_d q_\phi(\boldsymbol{z}_d)} - \ln\frac{p(\boldsymbol{z})}{\prod_d p(\boldsymbol{z}_d)}\right] - \\ &\quad \gamma \mathrm{E}_{q_\phi(\boldsymbol{z})}\left[\ln\frac{q_\phi(\boldsymbol{z}_d)}{\prod_e q_\phi(\boldsymbol{z}_{d,e})} - \ln\frac{p(\boldsymbol{z}_d)}{\prod_e p(\boldsymbol{z}_{d,e})}\right] - \\ &\quad \sum\nolimits_e \mathrm{KL}(q_\phi(\boldsymbol{z}_{d,e})||p(\boldsymbol{z}_{d,e})) \end{aligned} \tag{23}$$

3.3　树形网络

除了第 3.2 节中所展示的层次深度梯形网络架构的设计, 树形模型的结构设计更是将第 3.2 节中深度层次梯形网络与高层超潜变量间的横向连接思想相融合, 如图 7 所示, 构建出更符合现代神经科学在视觉皮层中观察到的横向连接现象. 将此结构归纳偏好再次加入模型结构设计中, 通过引入更深层的超潜变量父节点可以在达到子节点中潜在变量解耦效果的同时结合更深层父节点语义间的交互性特征, 实现更科学的解耦性能.

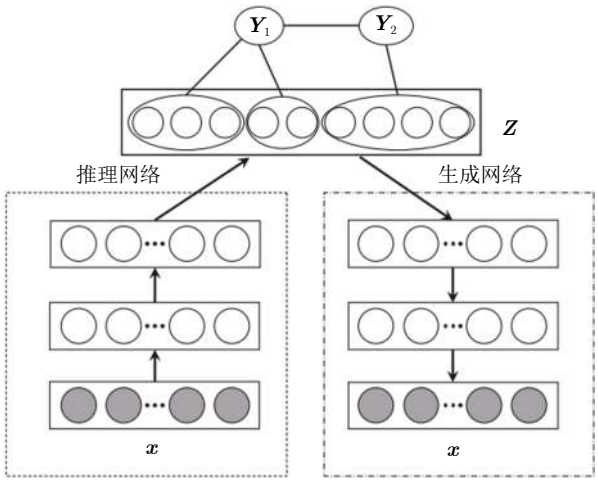


图 7　简易树形变分自编码模型示意图[73]

Fig. 7　Structure of a simple latent tree variational auto-encoders[73]

George 等[74] 于 2017 年所提出的递归皮层网络 (Recursive cortical network, RCN) 便搭建出一种类似人类大脑皮层处理方式的组合性树形结构网络, 如图 8 (d) 所示. 该网络将目标对象解耦为轮廓表征与外观表征, 如图 8 (a) 所示, 使模型能够识别具有明显不同外观的对象形状, 在复杂场景下的文本识别中展示了出色的泛化和遮挡推理的能力. 其中外观表征使用条件随机场 (Conditional random field, CRF) 建模来反映外观表面平滑变化的物理特性. 轮廓表征的建模方式如图 8 (b) 所示, 通过多层特征池、横向连接、组合的设计, 实现高层次特征间相互独立, 又同时共享底层特征彼此交互的特性. 其中池化结构的设计使得顶层特征节点能够表示具有一定平移、缩放和变形不变性的对象; 横向连接的设计能够实现同一层次不同组特征间的彼此交互作用, 其直观展示如图 8 (c) 所示. 该网络设计为组合式模型提供了更多的概率图模型中所涉及的高级推理与学习算法.

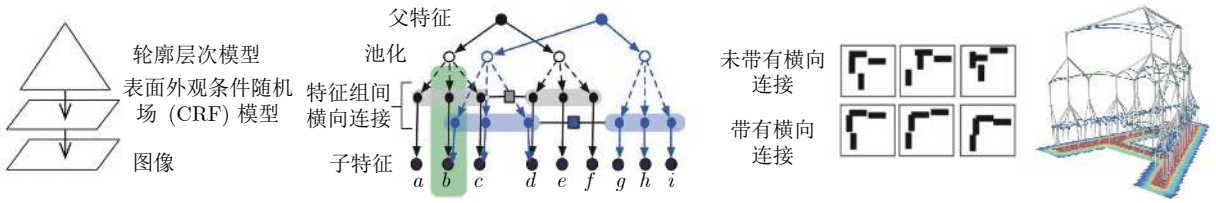
Li 等[73] 于 2019 年提出潜在树形变分自编码器 (Latent tree variational auto-encoders, LTVAE), 其表示结构是由多个超潜变量组成的树结构, 与 Willetts 等[71] 类似旨在生成多种按照数据不同指标方式的聚类结果. 该模型假设数据是通过神经网络从潜在特征生成的, 而潜在特征本身被另一层次的超潜变量通过树型贝叶斯网络生成, 每个超潜变量都代表着一种聚类方式. 该方法能够自主选择每个超潜变量的潜在特征子集, 并学习不同超潜变量间的依赖结构.

4　基于物理知识归纳偏好的解耦表征学习

除了第 2 和 3 节基于非结构化与结构化先验归纳偏好的解耦表征学习研究外, 在模型中融入真实数据内所蕴含的物理本征机理和复杂逻辑关系将有助于进一步发展内嵌底层逻辑与物理内涵的解耦表征学习新体系. 因此本节将从输入数据间的物理关联与基于对象的场景空间组合两种物理语义理解层面着手, 研究当下融入物理知识归纳偏好的解耦表征学习.

4.1　基于输入数据间物理关联的多输入解耦表征学习

第 2 和 3 节所述有关解耦表征学习的研究均默



(a) 轮廓表征与外观表征的层次化解耦结构示意图
(a) The hierarchy structure of disentangling the contour and appearance representation

(b) 包含池化结构与横向连接设计的轮廓层次结构示意图
(b) The structure of contour hierarchy including feature pooling and lateral connections

(c) 带有和不带有横向连接设计在一个文字角点处的直观展示图
(c) Samples from a corner feature with and without lateral connections

(d) 字母 “A” 的四级轮廓层次网络结构图
(d) Four-level network representing an “A”

图 8　RCN 模型示意图[74]

Fig. 8　Structure of the RCN[74]

认输入数据服从独立同分布特性, 然而在弱监督分组观测的情况下, 组内数据间蕴含着一定的共性特征(如各种视角下的同一物体共享身份特征; 同一颜色下的不同物体共享颜色特征等), 如图 9 中遥感舰船图像组数据示例, 此时组内数据间具有一定的相关性, 该假设不再成立. 因此本节旨在研究如何将组内数据间的弱监督相关性信息加入到网络归纳偏好的设计中, 实现针对分组数据的相关因子与不相关因子的解耦表征学习. 除此之外, 本节也会涉及到成对输入图像中感兴趣因子的差异比较, 这种比较与医生根据两个病人的疾病对比程度来量化他们的疾病严重程度思想类似, 旨在能够对相关感兴趣因子进行更好的量化.

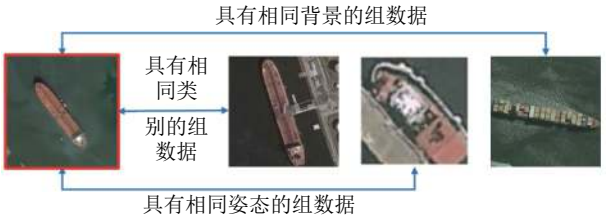


图 9 遥感舰船图像组数据示例图

Fig. 9 Samples from remote sensing ship group images

针对具有部分完全相同属性的分组数据而言, 大量研究学者提出通过在组内数据间共享或交换部分潜在变量的方法[75−79], 促使网络学习到代表组内数据间特定相关生成因子所对应的潜在变量, 通过该举措能够从施加强结构偏好的角度有效完成组内数据相关因子与不相关因子的解耦表征学习任务.

Bouchacourt 等于 2018 年[75] 提出多级变分自编码器 (Multi-level variational auto-encoders, ML-VAE), 在组内数据共享相关因子潜在表示 $\boldsymbol{C}_G$ 的同时, 认为不相关因子的潜在表示 $\boldsymbol{S}_G$ 服从独立阶乘分布, 二者共同参与图像的生成过程. 其中值得注意的是, 该架构构建了两组完全分离的编码网络, ϕ_c, ϕ_s 分别为这两组分离的编码网络的变分参数, 对两组隐变量语义表示进行源头性的阻断隔离解耦, 同时可以通过交换潜在表示生成新的类型图像对解耦表征进行可视化展示. 其整体的优化函数如式 (24) 所示.

$$\begin{aligned}&\mathcal{L}(\boldsymbol{G};\theta,\phi_c,\phi_s)=\\&\sum_{i\in G}\mathrm{E}_{q(\boldsymbol{C}_G|\boldsymbol{X}_G;\phi_c)}\left[\mathrm{E}_{q(\boldsymbol{S}_i|\boldsymbol{X}_i;\phi_s)}\left[\ln\left(\boldsymbol{X}_i|\boldsymbol{C}_G,\boldsymbol{S}_G;\theta\right)\right]\right]-\\&\sum_{i\in G}\mathrm{KL}\left(q\left(\boldsymbol{S}_i|\boldsymbol{X}_i;\phi_s\right)||p\left(\boldsymbol{S}_i\right)\right)-\\&\mathrm{KL}\left(q\left(\boldsymbol{C}_G|\boldsymbol{X}_G;\phi_c\right)||p\left(\boldsymbol{C}_G\right)\right)\end{aligned}\tag{24}$$

与 ML-VAE 组内数据共享相关因子的表示不同, Szabó等[77] 于 2018 年提出组内图像对间通过交换不相关因子的潜在变量表示来重构出其所对应的另一幅图像, 而不是原图像本身, 通过这种操作达到解耦相关因子与不相关因子的目的. 除此之外, 他们提出对于分组数据的解耦表征学习往往存在退化映射问题, 即图像的所有信息均集中在某一部分的特征表示中. 为了缓解这一问题, 他们引入与输入图像对 $\{x_1, x_2\}$ 均完全无关的图像 x_3, 运用对抗思想再次将潜在变量进行交换来保证相关因子的潜在变量表示一定包含部分图像信息, 有效避免了退化映射现象的发生. Ge 等[78] 于 2021 年提出组监督学习模型 GSL (Group-supervised learning), 在结合上述交换潜在变量表示进行回归匹配思想的同时融合 Cycle-GAN[80] 的思想, 将交换隐变量表示后的图像再次通过同样的网络将其交换回来进行与原图像间的回归匹配. 该思想相较于上述方法优势在于保证上述方法性能的同时, 通过再次交换回传可以进一步施加原转换问题的逆约束, 使得属性值不一样的相关因子 (如都含有颜色相关属性但属性值不同的组图像) 的解耦进一步明朗化. 以图 9 中的遥感舰船图像组数据为例, 若使用 GSL 模型, 则该模型对应的网络设计如图 10 所示.

除了上述通过共享或交换潜在变量来达到相关属性与其余属性的解耦表征外, 从互信息相关性角度对分组数据内潜在表示间进行相关性度量, 也可以进一步对分组数据输入施加正则约束, 从而促进分组数据潜在表示的挖掘与解耦. Sanchez 等[81] 便于 2020 年采用局部互信息与全局互信息相结合的方式衡量图像对内相关因子潜在表示的相关性, 让其值尽可能大, 促使分组数据间不同数据的相同属性表示尽可能相似. 同时为了达到解耦目的, 运用对抗思想来使同一数据内共享表示和互斥表示间的互信息尽可能小.

将上述共享或交换隐变量表示与互相关信息的思想相结合, Esser 等[82] 构建分离的姿态编码器与外观编码器, 从目标姿态所对应的图像中学习姿态表示, 从目标外观所对应的图像中学习外观表示, 随后共同送入解码网络中生成新的图像. 该网络在训练过程时与 Sanchez 等[81] 类似也采用判别器约束外观表示与姿态表示间的互信息大小. Lorenz 等[83] 在 Esser 等[82] 的基础上, 将前景目标看作由一系列部件通过一定的空间组合规律组成, 每个部件都具有外观与姿态特性, 除此之外他们还利用物理变换的方式人为将一幅图像扩充为姿态发生变化但外观未变的图像与外观未发生变化但姿态发生变化的图

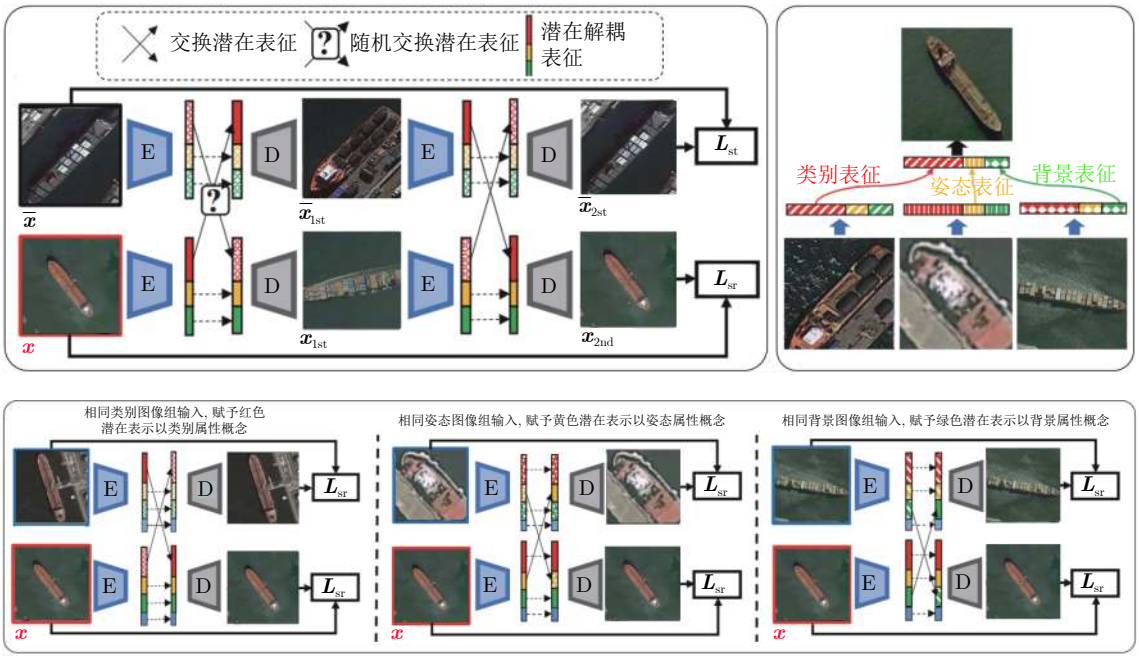


图 10　GSL 模型[78] 用在遥感舰船图像组数据集中对应的网络架构示意图

Fig. 10　The structure of GSL model[78] when it is used in the remote sensing ship image group data set

像来取代组标签信息, 从而设计分离的编码网络, 从外观变化的图像中学习姿态信息, 从姿态变化的图像中学习外观信息. 该网络设计能够在无监督条件下利用自监督思想有效实现部件间姿态与外观的解耦表征, 将分组数据间的解耦表征研究思想应用到通过数据增强等有效物理转换方式的独立数据解耦表征研究中. 此后这种通过物理变换构造分组数据以及姿态, 外观的解耦方式还被 Liu 等[84] 用于无监督部件分割的任务研究中. Dundar 等[85] 则是将上述方法扩充到视频信息中随时空变化与随时空不变的信息间的解耦表征. 他们认为相邻帧中除了背景信息随时间推移稳定不变外, 前景纹理信息在前景目标还未消失前也同样保持不变, 随时间变化的仅为前景目标的形态姿势信息. 基于此, 他们利用相邻帧之间前景目标姿态信息各异而外观信息与背景信息共享这一组内的弱监督信息出发, 创建出一种新颖的模型架构旨在将视频帧中前景与背景分离, 且前景信息中姿态信息与外观信息分离.

以上研究都是基于相同的指标属性进行分组解耦表征, 然而现实数据集中大多纷杂错乱, 如何综合利用按照各种不同指标的分组数据变成了一个新的挑战. Vowels 等[86] 于 2020 年提出了 Gated-VAE, 期望在网络训练过程中能够加入任何可用领域的先验知识, 使得模型的适用性更广. 他们提出一种新颖的训练方式, 在梯度前向传播过程中, 所有潜在空间的分区共同合并在一起, 即在整个潜在空间上进行优化函数的计算; 但在误差反向传播过程中, 梯度将根据不同的图像对来选择特定的潜在空间分区进行传播. 通过这种独特的训练方法, 他们认为如果整个数据集中所需的分区与人为所划分的图像对一致, 则各个分区将包含不同的因素. 即使在分区内存在纠缠现象, 分区间也会实现解耦, 为解耦学习的研究注入了新思路.

4.2　基于对象空间组合归纳偏好的解耦表征学习

正如第 3.1 节所述, 场景图像可看作由背景元素和以空间特定方式排列的多个离散对象组成, 而单个对象又可以看作是由外观与形状耦合而成. 因此本节注重于考虑如何将一幅复杂场景解构为多个简单对象的组合, 并据此理解/生成这些简单对象的组合关系.

人类天生具有组合泛化的能力, 如图 11 所示, 对于一幅多物体复杂场景, 人类可以将其解构为多个简单对象, 并可以在脑海中按照空间位置重新排列组合这些对象, 构成一幅新的场景图. 除此之外, 对于单个简单对象而言, 人类也可以将其解构为具有多组共通属性与各异属性的多个简单部件. 正是依靠这种组合泛化能力, 人类智能才能够从一些最基础的元素出发, 一步一步创造出复杂甚至无限的语义世界. 从这个角度出发, 越来越多的研究工作开始研究探索数据集中内在的组合性规律, 旨在促

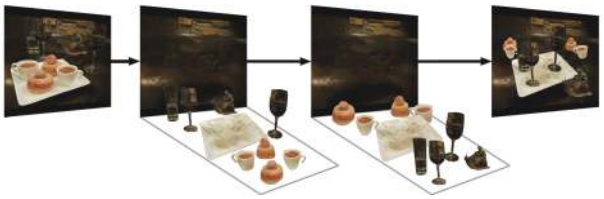


图 11 人类想象泛化能力示意图[87]

Fig. 11 An example of human imagination generalization ability[87]

使深度神经网络拥有像人类一样的组合泛化能力.

要拥有像人类一样的组合泛化能力, 首先应学会对各类输入数据进行内在分组解耦, 例如在解决鸡尾酒会的问题时, 应对不同说话人的语音进行解耦; 在自动驾驶中, 应对道路上各种不同对象的类别、位置和速度进行解耦. 在现实世界中, 这些信息或多或少相互纠缠, 隐藏在可见数据背后, 本小节将着重于解耦隐藏在真实数据背后的丰富物理结构, 完成不同对象不同特性的解耦表征任务.

Greff 等[88] 于 2016 年提出一种能够有效进行分组解耦的迭代推理框架 (TAGGER). 该框架对数据输入类型不设约束, 结合期望最大化 (Expectation-maximum, EM) 聚类算法, 对数据背后的潜在分组以及每个分组所对应的潜在表示进行迭代摊销推理: 给定分组的条件下推断各个组内特征; 给定各个组内特征的条件下推断分组, 如此迭代优化地完成对组分配以及各个组内对象表征的估计任务. 然而正是由于该方法对数据类型以及网络设计不施加任何其他约束, 仅能够对存在明显分组偏差的简单数据集进行解耦表征, 并未泛化到各类复杂任务场景中.

对于单目标场景的组合泛化能力而言, Li 等[89] 于 2020 年提出条件生成模型 MixNMatch (Mix-and-match), 旨在将单目标场景解耦为背景、前景目标的形状、姿态、外观四类表征. 他们对于场景生成过程的理解借鉴 Singh 等[90] 的 FineGAN 模型, 分离为三个独立作用的阶段: 背景提取阶段、父场景前景形状提取阶段和子场景外观提取阶段. 其中背景提取阶段利用控制背景图像生成的潜在变量进行背景建模. 父场景前景形状提取阶段利用控制对象的轮廓 (形状) 的潜在变量生成前景目标的形状掩模. 子场景外观提取阶段则利用控制对象外观纹理的潜在变量进行前景外观建模. 三个独立的生成网络间首尾相连将每个网络生成的图像拼接耦合在一起生成最终的细粒度图像. 尽管这类模型在单域数据集下能够达到很好的解耦表征性能, 具有一定的联想组合泛化能力, 但对于多个跨域数据集的解耦性能却不尽如人意. Ojha 等[91] 认为其原因在于源域属性信息在单域数据集下并未多做考虑, 导致其耦合在形状、外观等表征中. 对此, 他们在 FineGAN 的基础上提出一种基于可学习的物体外观特征直方图表示, 从而消除跨域情况下域信息对于物体解耦表征的影响.

除了将前景目标视为一个整体解耦其姿态与外观纹理属性外, Lorenz 等[83] 认为前景目标可解耦为不同部件的外观表示与姿态表示. 如此则应保证原部件施加外观转换干扰时其所对应的姿态表示不应发生变化, 反之亦然. 他们便将这种真实世界存在的物理约束加入模型设计中, 从部件外观变化的图像中学习部件姿态表示, 从部件姿态变化的图像中学习部件外观表示这种强逻辑结构. 但他们并未进一步考虑部件与整体间的逻辑映射关系. 对于此类问题, Kosiorek 等[92] 于 2019 年提出的堆栈胶囊自编码网络 (Stacked capsule auto-encoders, SCAE) 则巧妙运用自然语言处理领域内 Set transformer[93] 思想, 将部件组成整体的任意组合方式考虑进去. 该模型首先将图像分割为多个部件, 再将部件组合为多个连贯的整体, 整个逻辑图如图 12 所示, 不仅能够解决单目标场景的部件解耦问题, 还能够泛化到多目标场景的目标级解耦以及每个目标多对应的多部件解耦中, 为解耦问题注入了新思想. 此外, Yang 等[94] 认为解耦后的潜在变量在通过解码网络生成原图像的过程应服从一定的因果关系, 他们提出的 CausalVAE 在网络解码过程中加入了一层用于挖掘潜在变量间因果关系的因果层, 促使整个网络的生成过程更服从人类对于世界因果关系的认知过程, 为构建因果结构化的解耦表征学习模型提供了重大的参考意义.

针对多目标场景的对象级解耦表征理解, 除了上述的 SCAE 模型外, 2019 年, Greff 等[95] 提出的迭代对象分解推理网络 (Iterative object decomposition inference network, IODINE) 与 Bur-

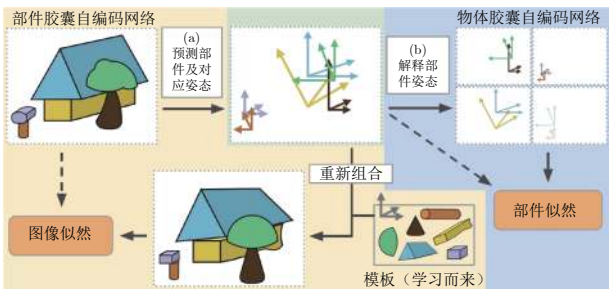


图 12 堆栈胶囊自编码网络 (SCAE) 模型架构图[92]

Fig. 12 Architecture of stacked capsule autoencoders (SCAE)[92]

gess 等[96] 提出多对象网络 (Multi-object network, MONET) 均将多目标场景图像理解为由多个物体级别的抽象块按照一定的空间映射关系组合而来. 基于此假设, 二者均将多目标场景图像分布视为由多个服从单高斯分布的物体级抽象块按照一定的概率组合而成的混合高斯分布. 其中对于推理网络的设计, IODINE 采用迭代变分推理的方式[97] 得到每个物体级抽象块所对应的潜在表征, 随后利用解码网络得到每个物体级抽象块所对应的高斯分布似然图以及空间掩码概率图. 而 MONET 则使用递归空间注意力网络得到每个物体级的抽象块所对应的空间掩码概率图, 随后将该概率图与原图一起输入自编码网络中得到每个物体级抽象块的高斯分布似然图. 然而这两个网络仅能处理简单多目标场景, 并不能解决复杂多目标场景的目标级解耦表征.

针对复杂多目标场景的解耦表征理解, Zhan 等[87] 于 2020 年提出了一种自监督的场景遮掩算法, 用于学习物体间相互遮挡的空间关系. 如图 13 所示, 该算法从前景目标间的空间排列组合方式提出了构建有向图进行表征的新角度, 通过目标间的空间逻辑树状图的构建实现了对多目标场景图像的空间想象.

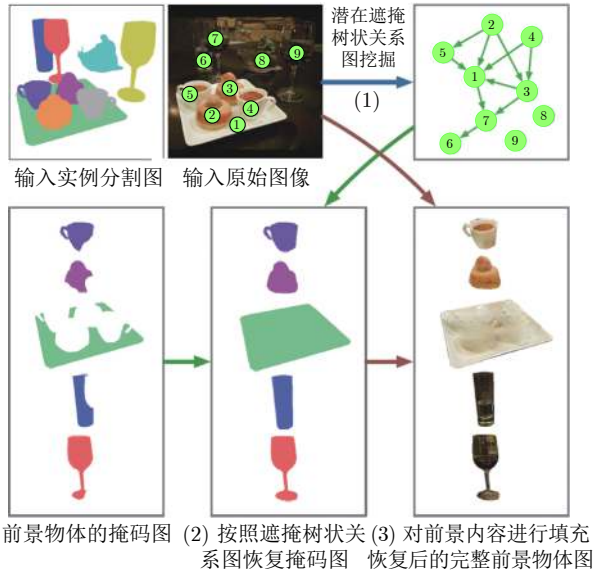


图 13　多目标场景去遮掩实现过程示意图[87]

Fig. 13　The framework of the de-occlusion completion for multi-objective scene[87]

除此之外, Prabhudesai 等[98] 于 2021 年提出从 3D 特征图的角度进行二维图像解耦的新思想. 他们认为对于一个前景目标而言, 三维本征立体结构是目标本身的内禀不变属性, 据此他们提出将二维图像投影到三维空间中, 在三维空间中进行对象级别的解耦更加能够符合人类对于二维图像数据的认知过程, 且通过此方式不仅能够解决目标遮掩与视角大差异性问题, 而且能够从任意视角想象泛化出新场景图像, 为真正从三维角度看待二维图像解耦表征问题提供了良好的解决思路.

5　模型对比分析

上文描述的有关解耦表征学习算法被归纳为三类: 基于非结构化表征先验的解耦表征学习、基于结构化模型先验归纳偏好的解耦表征学习与基于物理知识归纳偏好的解耦表征学习. 本节将对这三类方法进行对比分析, 讨论其各自的适用范围, 并选取部分模型进行实验性能的可视化展示, 突出解耦表征学习对各类下游任务以及可解释性深度学习的贡献.

解耦表征学习的真正内核在于将数据内部各个具有可解释性的生成因子采用尽可能独立的潜在变量子集进行捕获表征, 并不拘泥于特定的数据类型与具体的下游任务. 从表 2 中可以看出, 本文各类算法的适用场景以及对应的下游任务不尽相同, 所选用的数据集也各有侧重. 因此对于解耦表征学习而言, 各类算法的评价指标并未有统一标准, 应根据实际情况具体分析所需解决问题的特定数据集, 从主观的角度出发设定能为人类所理解的特定评价指标. 基于此, 本节仅选取几类典型算法的特定测试性能进行展示以辅助读者结合表 2 内容对解耦表征学习进一步深入思考.

首先针对第一大类基于非结构化表征先验的解耦表征算法而言, 该类算法大都属于无监督学习范畴, 通过在网络优化过程中对潜在表征施加独立性、稀疏性等归纳偏好, 约束网络学习可分离的潜在表征. 对于这些可分离表征的物理意义验证, 大多文献采用控制变量重构法进行直观可视化验证, 即在保持其余潜在表征不变的情况下, 依次单独变换某一特定潜在表征的取值, 通过分析重构图像所发生的可视化改变来对这一特定潜在表征所代表的具体物理含义进行人为分析理解, 如图 14 所示. 除此之外, Higgins 等[46], Kim 等[51], Kumar 等[49], Chen 等[52], Eastwood 等[130] 先后提出一系列有关潜在表征解耦性能的度量方法, 然而这些方法只是单纯探究潜在表征的可分离性, 对其背后所捕捉的可解释性因子并未进行评判. 而解耦表征学习的本质重在挖掘数据生成背后复杂耦合的物理机理, 并非单纯地学习一堆未知含义但拥有独立统计特性的潜在表征集合. 因此本节并未同 Locatello 等[37] 一样根据这些独立性指标对该类算法进行量化对比, 只是单纯列举其中典型算法 AAE[48], Factor-VAE[51] 在一些数字、人

表 2 不同归纳偏好方法对比

Table 2 Comparisons of methods based on different inductive bias

归纳偏好分类	模型	简要描述	适用范围	数据集
非结构化表征先验	β-VAE[46] InfoGAN[55] 文献 [47] Joint-VAE[53] AAE[48] DIP-VAE[49] Factor-VAE[51] RF-VAE[56] β-TCVAE[52]	在网络优化过程中施加表1中不同的先验正则项, 能够促使网络学习到的潜在表征具备一定的解耦性能. 但该类方法并未涉及足够的显式物理语义约束, 网络不一定按照人类理解的方式进行解耦, 因此该类方法一般用于规律性较强的简易数据集中.	适用于解耦表征存在显著可分离属性的简易数据集, 如人脸数据集、数字数据集等.	MNIST[99]; SVHN[100]; CelebA[101]; 2D Shapes[102]; 3D Chairs[103]; dSprites[102]; 3D Faces[104]
结构化模型先验	顺序深度递归网络 DRAW[62] AIR[64] SQAIR[66]	通过构建顺序深度递归网络架构, 可以在执行决策时反复结合历史状态特征, 实现如简易场景下的检测、跟踪等.	适用于需要关联记忆的多次决策任务场景.	3D scenes[64]; Multi-MNIST[64]; dSprites[102]; Moving-MNIST[66]; Omniglot[105]; Pedestrian CCTV data[106]
	层次深度梯形网络 VLAE[70] 文献 [71] HFVAE[72]	使用层次梯形网络模拟人类由浅入深的层次化认知过程, 促使每层潜在变量代表着不同的涵义, 可用作聚类等任务.	适用于简易数据集下由浅入深的属性挖掘.	MNIST[99]; CelebA[101]; SVHN[100]; dSprites[102]
	树形网络 RCN[74] LTVAE[73]	使用树形网络模拟人类高级神经元间的横向交互过程, 完成底层特征解耦的同时高层特征语义交互, 可用作聚类、自然场景文本识别等任务.	适用于底层特征解耦共享, 高级特征耦合交互的场景任务.	CAPTCHA[107]; ICDAR-13 Robust Reading[107]; MNIST[99]; HHAR[73]; Reuters[108]; STL-10[73]
	分组数据的相关性 MLVAE[75] 文献 [77] GSL[78] 文献 [81] 文献 [82] 文献 [83] 文献 [85] 文献 [86]	通过交换、共享潜在表征、限制互信息相关性、循环回归等方式, 实现分组数据相关因子的解耦表征. 后续可单独利用有效因子表征实现分类、分割、属性迁移数据集生成等任务.	适用于分组数据的相关有效属性挖掘.	MNIST[99]; RaFD[109]; Fonts[78]; CelebA[101]; Colored-MNIST[81]; dSprites[102]; MS-Celeb-1M[110]; CUB birds[111]; ShapeNet[112]; iLab-20M[113]; 3D Shapes[81]; IAM[114]; PKU vehicle id[115]; Sentinel-2[116]; Norb[117]; BBC Pose dataset[118]; NTU[119]; KTH[120]; Deep fashion[121]; Cat head[122]; Human3.6M[123]; Penn action[124]; 3D cars[125]
物理知识先验	基于对象的物理空间组合关系 MixNMatch[89]	结合数据组件化、层次化生成过程实现单目标场景的背景、姿态、纹理、形状解耦表征.	适用于单目标场景属性迁移的数据集生成.	CUB birds[111]; Stanford dogs[126]; Stanford cars[125]
	文献 [83]	考虑单目标多部件间的组合关系.	适用于人类特定部位、面部表情转换等数据生成.	Cat head[122]; Human 3.6M[123]; Penn action[124]
	SCAE[92]	提出了胶囊网络的新思想, 考虑多目标、多部件间的组合关联关系.	适用于简易数据集的目标、部件挖掘.	MNIST[99]; SVHN[100]; CIFAR10
	TAGGER[88] IODINE[95] MONET[96]	考虑多目标场景的逐次单目标解耦表征方式.	适用于简易多目标场景的目标自主解译任务.	Shapes[127]; Textured MNIST[88]; CLEVR[128]; dSprites[102]; Tetris[95]; Objects room[96]
	文献 [87]	引入目标空间逻辑树状图, 解耦多目标复杂场景的遮掩关系, 可用于去遮挡等任务.	适用于自然复杂场景下少量目标的去遮挡任务.	KINS[129]; COCOA[112]
	文献 [98]	将目标三维本体特征视为目标内禀不变属性进行挖掘, 解决视角、尺度大差异问题, 有望实现检测、识别、智能问答等高级场景理解任务.	适用于简易数据集的高级场景理解.	CLEVR[128]

脸等简易数据集上的可视化解耦表征结果, 分别如图 14、15 所示.

其次针对基于结构化模型先验归纳偏好的解耦表征学习算法而言, 这类算法的独特之处在于模仿人类大脑的功能性区域构建可解释的网络架构. 这类由网络架构引起的模型结构化归纳偏好能够在很大程度上调整网络学习的方式, 其中顺序递归网络架构促使网络每做出一次决策的时候都会与之前学习的内容进行关联; 深度梯形网络架构促使网络由浅入深地逐层挖掘数据特征; 树形网络架构则会促使网络对高层高级特征进行横向语义关联. 因此这类由人类大脑结构启发的网络架构设计形式不尽相同, 所对应的人为任务偏好也千差万别. 为了形象化展示这类算法在解耦表征以及下游任务中所展现

图 14　Factor-VAE[51] 算法在 3D chairs[103] 以及 3D faces[104] 数据集上的解耦性能展示图. 每一行代表仅有左侧标注的潜在表征取值发生改变时所对应的重构图像变化

Fig. 14　The disentangled performance of Factor-VAE[51] for 3D chairs[103] and 3D faces[104] data sets. Each row represents the change in the image reconstruction when only the specific latent marked on the left change

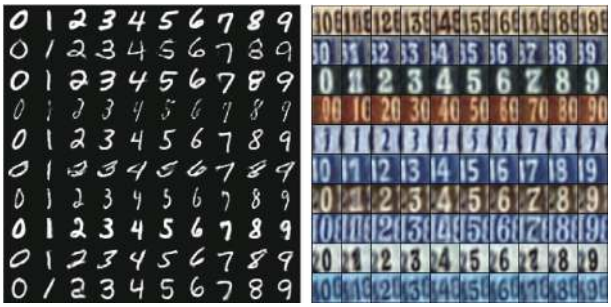


图 15　AAE[48] 算法对于 MNIST[99] 和 SVHN[100] 数字数据集中类别与风格属性的解耦表征结果展示图. 图中每一行代表风格类潜在表征保持不变的情况下, 改变类别类潜在表征取值所对应的重构图像变化; 每一列代表类别类潜在表征保持不变的情况下, 改变风格类潜在表征取值所对应的重构图像变化

Fig. 15　The disentangled performance of AAE[48] in the MNIST[99] and SVHN[100] data set. Each row represents the change of the reconstructed images corresponding to the category latent while the style latent remains unchanged; when each column represents the change of the reconstructed images corresponding to the style latent while the category latent is unchanged

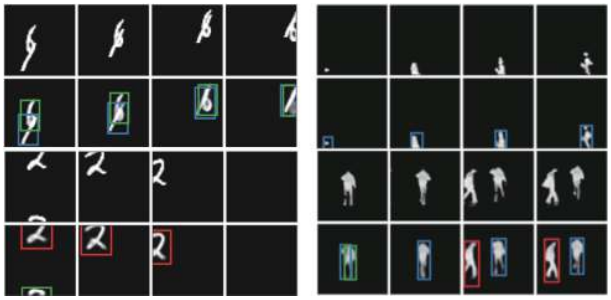


图 16　SQAIR[66] 用于视频目标检测、跟踪实验结果图. 其中不同颜色的标注框代表网络递归过程中所检测、跟踪到的不同目标

Fig. 16　The video target detection and tracking results of SQAIR[66], where the bounding boxes with different colors represent different objects

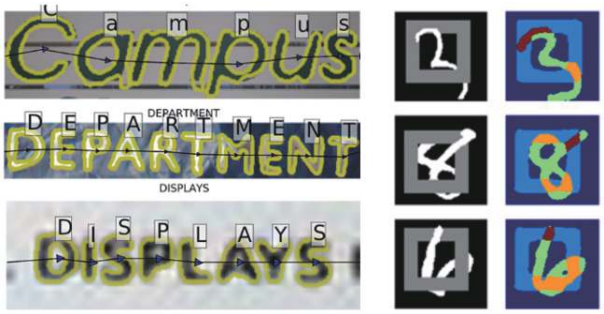


图 17　RCN[74] 用于字符分割识别的实验结果展示图. 其中左侧图像中黄色轮廓线为字符分割结果, 右侧第一列为输入遮掩数字, 第二列为网络预测的去遮掩掩码图

Fig. 17　Scene-text parsing results with RCN[74]. The yellow outline in the left image shows segmentations, the first column on the right is the occlusion input, and the second column shows the predicted occlusion mask

的优秀性能, 本节挑选出三类典型的模型架构代表算法 SQAIR[66], RCN[74], LTVAE[73], 验证解耦表征学习对于一些下游检测、识别、聚类等任务的有效性, 如图 16 ~ 18 所示.

最后针对基于物理知识归纳偏好的解耦表征学习算法而言, 该类算法更是将前两类算法与真实世界的物理知识相结合, 进一步提高了解耦表征学习的科学性. 本文将目前已有的相关类研究算法分为分组弱相关物理知识与对象空间组合关系物理知识两类, 其中前者旨在利用弱监督组信息去挖掘组内数据相关性特征, 这类算法的直观可视化验证主要通过属性迁移图像生成来验证相关性特征提取的好坏, 对此本节以文献 GSL[78] 为例, 直观展示其在属性迁移图像生成中所展现出的实验性能, 如图 19 所示. 对于后者基于对象空间组合关系的物理知识运用而言, 现有文献主要从多目标间的组合关联关系、目标内多部件间的组合关联关系两层面入手进行研究, 因此本节以文献 [83] 为例展示目标部件间的解耦表征学习性能, 如图 20 所示, 以文献 [87] 与文献 [98] 为例展示多目标场景下目标关系的重组化以及需要高级语义理解所支撑的智能问答任务性能, 分别如图 21、22 所示.

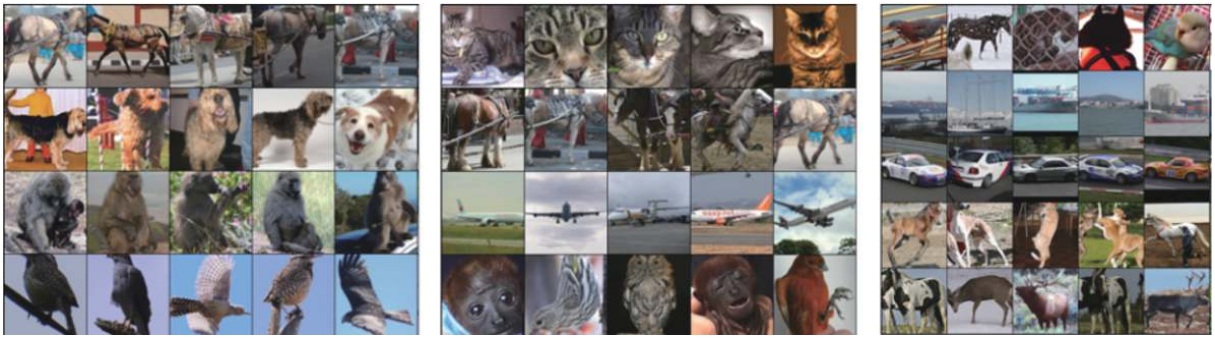


图 18 文献 [73] 所提算法的聚类实验结果图

Fig. 18 The clustering results of the algorithm proposed in the reference [73]

图 19 GSL[78] 算法所实现的图像属性迁移实验结果图

Fig. 19 The image synthesis qualitative performance by GSL[78]

6 研究重点和技术发展趋势

与人类相比, 目前的深度学习网络大多仅对与特定任务相关的样本特征进行训练, 而不考虑产生这些特征的内在物理属性, 因此在面对之前未见过的纠缠图像特征时表现出较弱的概括性与泛化性. 若深度网络能够学习到可概括的公共属性, 即带有实际物理语义属性的解耦表征学习, 将能够帮助神经网络想象各类具有不同属性的物体, 将已知耦合的图像特征转换为新的耦合图像特征, 例如, 可以对红船和蓝车的图像进行分解和重组, 合成新的红车图像等, 这将更有利于深度学习对数据内在本身物理特性进行挖掘, 增强对各类下游任务的迁移性与鲁棒性. 解耦表征的目的便是挖掘数据中潜在的相互作用因子, 并赋予其相互分离的数据表征, 属于可解释性的深度表征学习范畴, 能够很大程度上提高深度学习的可解释性, 增强其内在逻辑性, 在当今深度学习盛行的时代具有广阔的研究前景. 本文将目前有关解耦表征学习的研究大致概括

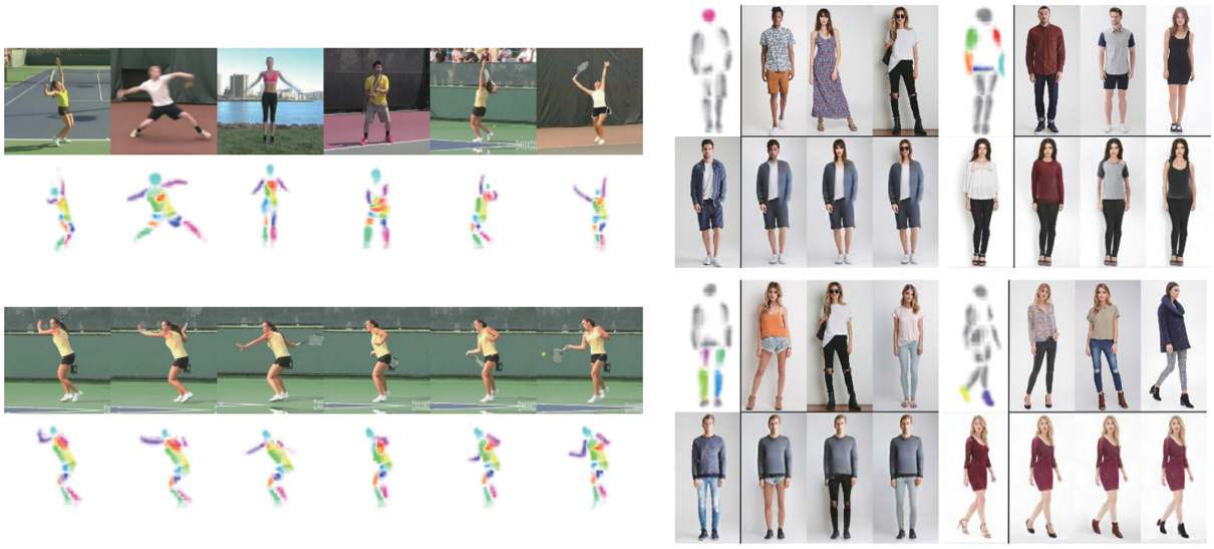


图 20 文献 [83] 所提算法在人类关节动作识别以及部分关节风格转换后生成图像的实验结果图

Fig. 20 The human action recognition and swapping part appearance results of the algorithm proposed in the reference [83]

图 21 文献 [87] 所提算法在自然场景下按照人类偏好重组目标位置以及遮盖顺序后的实验结果图

Fig. 21 The generation results of the algorithm proposed in the reference [87] after reorganizing the target position and the masking order in a natural scene

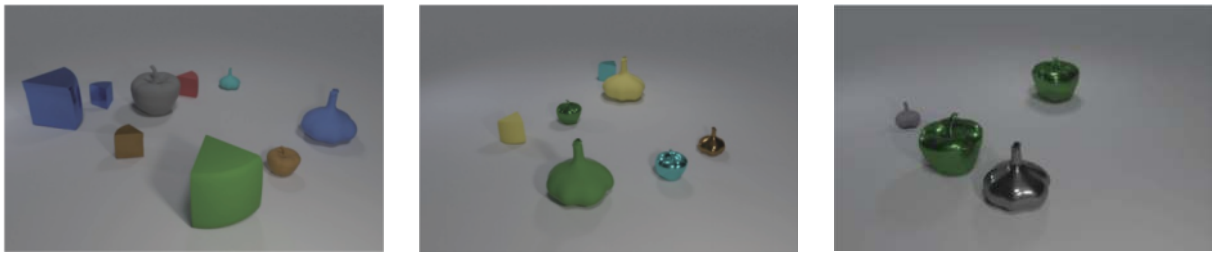


图 22 文献 [98] 所提方法应用在 CLEVR[128] 数据集上的智能知识问答实验结果图

Fig. 22 The VQA results on the CLEVR[128] data set using the method proposed in the reference [98]

为三类:

1) 基于非结构化表征先验的解耦表征学习, 如 β-VAE, InfoGAN, FactorVAE 等. 通过将潜在变量的先验分布的独立性约束传递给后验分布, 促使模型学习可分离的潜在变量表示, 从而达到解耦表征的效果. 然而该类方法并未考虑真实世界的复杂语义信息, 赋予潜在变量明确的物理含义, 导致其只能应用于手写数字体 MNIST 数据集、人脸 CelebA 数据集等简单数据集的解耦表征.

2) 基于结构化模型先验归纳偏好的解耦表征学习. 该类模型架构归纳偏差的设计主要基于类脑的思想构建深度层次化结构表征, 各个不同深度的层次代表不同语义信息, 如深度梯形网络、深度递归网络、树形网络等. 这类结构化归纳偏好的设计旨在挖掘自底向上、逐层递进的数据表征, 能够处理复杂场景大规模数据集以及数据流信息的解耦表征. 然而这类架构若仅仅模拟人脑结构, 并未赋予其更强的逻辑语义约束, 便不能真正达到符合人类理解的解耦表征学习.

3) 基于物理知识归纳偏好的解耦表征学习. 该类解耦表征学习旨在将强先验物理语义信息与逻辑关系加入模型设计中, 如多输入数据间的物理关联性、部件–个体间的逻辑拓扑关系, 个体–整体间的空间物理关系等, 能够同时融入上述两类归纳偏好的设计构成最终的解耦表征模型, 完成数据内部语义空间的挖掘, 能够处理复杂自然场景的数据.

本文对目前的解耦表征学习研究进行归纳总结后, 认为该研究领域依旧面临着许多严峻的挑战, 具有着广泛的研究前景. 以下是对该领域技术发展趋势的展望:

1) 建立世界的因果模型, 以支持解释和理解, 而不只是解决模式识别问题;

2) 物理和心理学的直观理论基础学习, 以支持和丰富所学习的知识;

3) 利用组合能力学习快速获取知识, 并将知识推广到新的任务和情况;

4) 提出能够量化由不同模型实现的解耦程度非常重要. 但是, 为此设计度量标准并不容易. 除了主观解释之外, 尽管有大量学者提出各种指标, 如分离属性可预测性[49]、互信息差异[52]、FactorVAE 度量[51]、β-VAE 度量[46]、解耦性/完整性/信息性 (Disentanglement/Completeness/Informativeness, DCI) 度量[130]、属性依赖关系 (Attribute dependency, AD) 度量[131] 等, 但目前还没有就定量衡量解耦性能的最佳标准达成共识, 这些指标中是否有任何一个能像人们通常想象的那样衡量解耦程度尚不清楚.

在当今深度学习快速发展的背景下, 泛化性与可解释性成为制约其进一步突破的关键问题, 受到社会各界的广泛关注. 解耦表征学习旨在挖掘数据内部潜在生成因子, 并利用可分离的潜在表示分别对其进行表征控制, 对数据进行深入理解, 揭示数据内部的生成作用机理, 逐渐成为提高深度学习泛化性、可扩展性与可解释性的重要手段. 本文对当前解耦表征学习研究进行了归纳总结, 该研究作为一门快速发展的开放性学科领域, 在内涵外延、模型理论、技术方法及实施策略方面还需要大量学者继续投入更多的研究与实践.

References

1 Duan Yan-Jie, Lv Yi-Sheng, Zhang Jie, Zhao Xue-Liang, Wang Fei-Yue. Deep learning for control: The state of the art and prospects. *Acta Automatica Sinica*, 2016, **42**(5): 643−654
(段艳杰, 吕宜生, 张杰, 赵学亮, 王飞跃. 深度学习在控制领域的研究现状与展望. 自动化学报, 2016, **42**(5): 643−654)

2 Wang Xiao-Feng, Yang Ya-Dong. Research on structure model of general intelligent system based on ecological evolution. *Acta Automatica Sinica*, 2020, **46**(5): 1017−1030
(王晓峰, 杨亚东. 基于生态演化的通用智能系统结构模型研究. 自动化学报, 2020, **46**(5): 1017−1030)

3 Amizadeh S, Palangi H, Polozov O, Huang Y C, Koishida K. Neuro-Symbolic visual reasoning: Disentangling "visual" from "reasoning". In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: PMLR, 2020. 279−290

4 Adel T, Zhao H, Turner R E. Continual learning with adaptive weights (CLAW). In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020.

5 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

6 Lee G, Li H Z. Modeling code-switch languages using bilingual parallel corpus. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020. 860−870

7 Chen X H. Simulation of English speech emotion recognition based on transfer learning and CNN neural network. *Journal of Intelligent & Fuzzy Systems*, 2021, **40**(2): 2349−2360

8 Lü Y, Lin H, Wu P P, Chen Y T. Feature compensation based on independent noise estimation for robust speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021, **2021**(1): Article No. 22

9 Torfi A, Shirvani R A, Keneshloo Y, Tavaf N, Fox E A. Natural language processing advancements by deep learning: A survey. [Online], available: https://arxiv.org/abs/2003.01200, February 27, 2020

10 Stoll S, Camgoz N C, Hadfield S, Bowden R. Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 2020, **128**(4): 891−908

11 He P C, Liu X D, Gao J F, Chen W Z. DeBERTa: Decoding-enhanced Bert with disentangled attention. In: Proceedings of the 9th International Conference on Learning Representations. Austria: ICLR, 2021.

12 Shi Y C, Yu X, Sohn K, Chandraker M, Jain A K. Towards universal representation learning for deep face recognition. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 6816−6825

13 Ni T G, Gu X Q, Zhang C, Wang W B, Fan Y Q. Multi-Task deep metric learning with boundary discriminative information for cross-age face verification. *Journal of Grid Computing*, 2020, **18**(2): 197−210

14 Shi X, Yang C X, Xia X, Chai X J. Deep cross-species feature learning for animal face recognition via residual interspecies equivariant network. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 667−682

15 Chen J T, Lei B W, Song Q Y, Ying H C, Chen D Z, Wu J. A hierarchical graph network for 3D object detection on point clouds. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 389−398

16 Jiang Hong-Yi, Wang Yong-Juan, Kang Jin-Yu. A survey of object detection models and its optimization methods. *Acta Automatica Sinica*, 2021, **47**(6): 1232−1255
(蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述. 自动化学报, 2021, **47**(6): 1232−1255)

17 Xu Z J, Hrustic E, Vivet D. CenterNet heatmap propagation for real-time video object detection. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 220−234

18 Zhang D W, Tian H B, Han J G. Few-cost salient object detection with adversarial-paced learning. [Online], available: https://arxiv.org/abs/2104.01928, April 5, 2021

19 Zhang Hui, Wang Kun-Feng, Wang Fei-Yue. Advances and perspectives on applications of deep learning in visual object detection. *Acta Automatica Sinica*, 2017, **43**(8): 1289−1305
(张慧, 王坤峰, 王飞跃. 深度学习在目标视觉检测中的应用进展与展望. 自动化学报, 2017, **43**(8): 1289−1305)

20 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

21 Geirhos R, Jacobsen J H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020, **2**(11): 665−673

22 Minderer M, Bachem O, Houlsby N, Tschannen M. Automatic shortcut removal for self-supervised representation learning. In: Proceedings of the 37th International Conference on Machine Learning. San Diego, USA: JMLR, 2020. 6927−6937

23 Ran X M, Xu M K, Mei L R, Xu Q, Liu Q Y. Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation. [Online], available: https://arxiv.org/abs/2007.08128v3, November 1, 2020

24 Charakorn R, Thawornwattana Y, Itthipuripat S, Pawlowski N, Manoonpong P, Dilokthanakul N. An explicit local and global representation disentanglement framework with applications in deep clustering and unsupervised object detection. [Online], available: https://arxiv.org/abs/2001.08957, February 24, 2020

25 Zhang Bo, Zhu Jun, Su Hang. Toward the third generation of artificial intelligence. *Scientia Sinica Informationis*, 2020, **50**(9): 1281−1302
(张钹, 朱军, 苏航. 迈向第三代人工智能. 中国科学: 信息科学, 2020, **50**(9): 1281−1302)

26 Lake B M, Ullman T D, Tenenbaum J B, Gershman S J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 2017, **40**: Article No. e253

27 Geirhos R, Meding K, Wichmann F A. Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. [Online], available: https://arxiv.org/abs/2006.16736v3, December 18, 2020

28 Regazzoni C S, Marcenaro L, Campo D, Rinner B. Multisensorial generative and descriptive self-awareness models for autonomous systems. *Proceedings of the IEEE*, 2020, **108**(7): 987−1010

29 Wang T, Huang J Q, Zhang H W, Sun Q R. Visual commonsense R-CNN. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 10757−10767

30 Wang T, Huang J Q, Zhang H W, Sun Q R. Visual commonsense representation learning via causal inference. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2020. 1547−1550

31 Schölkopf B, Locatello F, Bauer S, Ke N R, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. *Proceedings of the IEEE*, 2021, **109**(5): 612−634

32 Locatello F, Tschannen M, Bauer S, Rätsch G, Schölkopf B, Bachem O. Disentangling factors of variations using few labels. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020.

33 Dittadi A, Träuble F, Locatello F, Wüthrich M, Agrawal V, Winther O, et al. On the transfer of disentangled representations in realistic settings. In: Proceedings of the 9th International Conference on Learning Representations. Austria: ICLR, 2021.

34 Tschannen M, Bachem O, Lucic M. Recent advances in autoencoder-based representation learning. [Online], available: https://arxiv.org/abs/1812.05069, December 12, 2018

35 Shu R, Chen Y N, Kumar A, Ermon S, Poole B. Weakly supervised disentanglement with guarantees. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR, 2020.

36 Kim H, Shin S, Jang J, Song K, Joo W, Kang W, et al. Counterfactual fairness with disentangled causal effect variational autoencoder. In: Proceedings of the 35th Conference on Artificial Intelligence. Palo Alto, USA, 2021. 8128−8136

37 Locatello F, Bauer S, Lucic M, Rätsch G, Gelly S, Schölkopf B, et al. Challenging common assumptions in the unsupervised learning of disentangled representations. In: Proceedings of the 36th International Conference on Machine Learning. JMLR, 2019. 4114−4124

38 Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(8): 1798−1828

39 Sikka H. A Deeper Look at the unsupervised learning of disentangled representations in Beta-VAE from the perspective of core object recognition. [Online], available: https://arxiv.org/abs/2005.07114, April 25, 2020.

40 Locatello F, Poole B, Rätsch G, Schölkopf B, Bachem O, Tschannen M. Weakly-supervised disentanglement without compromises. In: Proceedings of the 37th International Conference on Machine Learning. San Diego, USA: JMLR, 2020. 6348−6359

41 Zhai Zheng-Li, Liang Zhen-Ming, Zhou Wei, Sun Xia. Research overview of variational auto-encoders models. *Computer Engineering and Applications*, 2019, **55**(3): 1−9
(翟正利, 梁振明, 周炜, 孙霞. 变分自编码器模型综述. 计算机工程与应用, 2019, **55**(3): 1−9)

42 Schmidhuber J. Learning factorial codes by predictability minimization. *Neural Computation*, 1992, **4**(6): 863−879

43 Kingma D P, Welling M. Auto-encoding variational Bayes. [Online], available: https://arxiv.org/abs/1312.6114, May 1, 2014

44 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS, 2014. 2672−2680

45 Lin Yi-Lun, Dai Xing-Yuan, Li Li, Wang Xiao, Wang Fei-Yue. The new frontier of AI research: Generative adversarial networks. *Acta Automatica Sinica*, 2018, **44**(5): 775−792
(林懿伦, 戴星原, 李力, 王晓, 王飞跃. 人工智能研究的新前线: 生成式对抗网络. 自动化学报, 2018, **44**(5): 775−792)

46 Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. Beta-vae: Learning basic visual concepts with a constrained variational framework. In: Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017.

47 Burgess C P, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, et al. Understanding disentangling in Beta-VAE. [Online], available: https://arxiv.org/abs/1804.03599, April 10, 2018

48 Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. [Online], available: https://arxiv.org/abs/1511.05644, May 25, 2016.

49 Kumar A, Sattigeri P, Balakrishnan A. Variational inference of disentangled latent concepts from unlabeled observations. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018.

50 Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. [Online], available: https://arxiv.org/abs/1701.04862, January 17, 2017

51 Kim H, Mnih A. Disentangling by factorising. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: JMLR, 2018. 2649−2658

52 Chen T Q, Li X C, Grosse R B, Duvenaud D. Isolating sources of disentanglement in variational autoencoders. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: NIPS, 2018. 2615−2625

53 Dupont E. Learning disentangled joint continuous and discrete representations. [Online], available: https://arxiv.org/abs/1804.00104v3, October 22, 2018.

54 Maddison C J, Mnih A, Teh Y W. The concrete distribution: A continuous relaxation of discrete random variables. [Online], available: https://arxiv.org/abs/1611.00712, March 5, 2017.

55 Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS, 2016. 2180−2188

56 Kim M, Wang Y T, Sahu P, Pavlovic V. Relevance factor VAE: Learning and identifying disentangled factors. [Online], available: https://arxiv.org/abs/1902.01568, February 5, 2019.

57 Grathwohl W, Wilson A. Disentangling space and time in video with hierarchical variational auto-encoders. [Online], available: https://arxiv.org/abs/1612.04440, December 19, 2016.

58 Kim M, Wang Y T, Sahu P, Pavlovic V. Bayes-factor-VAE: Hierarchical Bayesian deep auto-encoder models for factor disentanglement. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 2979−2987

59 Montero M L, Ludwig C J H, Costa R P, Malhotra G, Bowers J S. The role of disentanglement in generalisation. In: Proceedings of the 9th International Conference on Learning Representations. Austria: ICLR, 2021.

60 Larochelle H, Hinton G E. Learning to combine foveal glimpses with a third-order boltzmann machine. *Advances in Neural Information Processing Systems*, 2010, **23**: 1243−1251

61 Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS, 2014. 2204−2212

62 Gregor K, Danihelka I, Graves A, Rezende D J, Wierstra D. DRAW: A recurrent neural network for image generation. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR, 2015. 1462−1471

63 Henderson J M, Hollingworth A. High-level scene perception. *Annual Review of Psychology*, 1999, **50**(1): 243−271

64 Eslami S M A, Heess N, Weber T, Tassa Y, Szepesvari D, Kavukcuoglu K, et al. Attend, infer, repeat: Fast scene understanding with generative models. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: NIPS, 2016. 3233−3241

65 Crawford E, Pineau J. Spatially invariant unsupervised object detection with convolutional neural networks. In: Proceedings of the 33rd Conference on Artificial Intelligence. California, USA: AAAI, 2019. 3412−3420

66 Kosiorek A R, Kim H, Posner I, Teh Y W. Sequential attend, infer, repeat: Generative modelling of moving objects. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: NIPS, 2018. 8615−8625

67 Santoro A, Raposo D, Barrett D G T, Malinowski M, Pascanu R, Battaglia P W, et al. A simple neural network module for relational reasoning. In: Proceedings of the 31th International Conference on Neural Information Processing Systems. Long Beach, USA: NIPS, 2017. 4967−4976

68 Massague A C, Zhang C, Feric Z, Camps O I, Yu R. Learning disentangled representations of video with missing data. In: Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: California, USA, 2020. 3625−3635

69 Sønderby C K, Raiko T, Maaløe L, Sønderby S K, Winther O. Ladder variational autoencoders. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: NIPS, 2016. 3745−3753

70 Zhao S J, Song J M, Ermon S. Learning hierarchical features from deep generative models. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR, 2017. 4091−4099

71 Willetts M, Roberts S, Holmes C. Disentangling to cluster: Gaussian mixture variational Ladder autoencoders. [Online], available: https://arxiv.org/abs/1909.11501, December 4, 2019.

72 Esmaeili B, Wu H, Jain S, Bozkurt A, Siddharth N, Paige B, et al. Structured disentangled representations. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Okinawa, Japan: AISTATS, 2019. 2525−2534

73 Li X P, Chen Z R, Poon L K M, Zhang N L. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: ICLR, 2019.

74 George D, Lehrach W, Kansky K, Lázaro-Gredilla M, Laan C, Marthi B, et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science*, 2017, **358**(6368): eaag2612

75 Bouchacourt D, Tomioka R, Nowozin S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 2095−2102

76 Hwang H J, Kim G H, Hong S, Kim K E. Variational interaction information maximization for cross-domain disentanglement. In: Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: California, USA, 2020. 22479−22491

77 Szabó A, Hu Q Y, Portenier T, Zwicker M, Favaro P. Understanding degeneracies and ambiguities in attribute transfer. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 721−736

78 Ge Y H, Abu-El-Haija S, Xin G, Itti L. Zero-shot synthesis with group-supervised learning. In: Proceedings of the 9th International Conference on Learning Representations. Austria: ICLR, 2021.

79 Lee S, Cho S, Im S. DRANet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 15247−15256

80 Zhu J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2242−2251

81 Sanchez E H, Serrurier M, Ortner M. Learning disentangled representations via mutual information estimation. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 205−221

82 Esser P, Haux J, Ommer B. Unsupervised robust disentangling of latent characteristics for image synthesis. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 2699−2709

83 Lorenz D, Bereska L, Milbich T, Ommer B. Unsupervised part-based disentangling of object shape and appearance. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 10947−10956

84 Liu S L, Zhang L, Yang X, Su H, Zhu J. Unsupervised part segmentation through disentangling appearance and shape. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 8351−8360

85 Dundar A, Shih K, Garg A, Pottorff R, Tao A, Catanzaro B. Unsupervised disentanglement of pose, appearance and background from images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, DOI: 10.1109/TPAMI.2021.3055560

86 Vowels M J, Camgoz N C, Bowden R. Gated variational autoencoders: Incorporating weak supervision to encourage disen-

tanglement. In: Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). Buenos Aires, Argentina: IEEE, 2020. 125−132

87 Zhan X H, Pan X G, Dai B, Liu Z W, Lin D H, Loy C C. Self-supervised scene de-occlusion. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 3783−3791

88 Greff K, Rasmus A, Berglund M, Hao T H, Schmidhuber J, Valpola H. Tagger: Deep unsupervised perceptual grouping. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: NIPS, 2016. 4491−4499

89 Li Y H, Singh K K, Ojha U, Lee Y J. MixNMatch: Multifactor disentanglement and encoding for conditional image generation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 8036−8045

90 Singh K K, Ojha U, Lee Y J. FineGAN: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 6483−6492

91 Ojha U, Singh K K, Lee Y J. Generating furry cars: Disentangling object shape & Appearance across Multiple Domains. In: Proceedings of the 9th International Conference on Learning Representations. Austria: ICLR, 2021.

92 Kosiorek A R, Sabour S, Teh Y W, Hinton G E. Stacked capsule autoencoders. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: NIPS, 2019. 15512−15522

93 Lee J, Lee Y, Kim J, Kosiorek A R, Choi S, Teh Y W. Set transformer: A framework for attention-based permutation-invariant neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: JMLR, 2019. 3744−3753

94 Yang M Y, Liu F R, Chen Z T, Shen X W, Hao J Y, Wang J. CausalVAE: Disentangled representation learning via neural structural causal models. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 9588−9597

95 Greff K, Kaufman R L, Kabra R, Watters N, Burgess C, Zoran D, et al. Multi-object representation learning with iterative variational inference. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: JMLR, 2019. 2424−2433

96 Burgess C P, Matthey L, Watters N, Kabra R, Higgins I, Botvinick M, et al. MONet: Unsupervised scene decomposition and representation. [Online], available: https://arxiv.org/abs/1901.11390, January 22, 2019

97 Marino J, Yue Y, Mandt S. Iterative amortized inference. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: JMLR, 2018. 3400−3409

98 Prabhudesai M, Lal S, Patil D, Tung H Y, Harley A W, Fragkiadaki K. Disentangling 3D prototypical networks for few-shot concept learning. [Online], available: https://arxiv.org/abs/2011.03367, July 20, 2021

99 Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

100 Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng A Y. Reading digits in natural images with unsupervised feature learning. In: Proceedings of Advances in Neural Information Processing Systems. Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain: NIPS, 2011. 1−9

101 Liu Z W, Luo P, Wang X G, Tang X O. Deep learning face attributes in the wild. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 3730−3738

102 Matthey L, Higgins I, Hassabis D, Lerchner A. dSprites: Disentanglement testing sprites dataset [Online], available: https://github.com/deepmind/dsprites-dataset, Jun 2, 2017

103 Aubry M, Maturana D, Efros A A, Russell B C, Sivic J. Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 3762−3769

104 Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T. A 3D face model for pose and illumination invariant face recognition. In: Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance. Genova, Italy: IEEE, 2009. 296−301

105 Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction. *Science*, 2015, **350**(6266): 1332−1338

106 Ristani E, Solera F, Zou R S, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 17−35

107 Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda L G I, Mestre S R, et al. ICDAR 2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, USA: IEEE, 2013. 1484−1493

108 Xie J Y, Girshick R B, Farhadi A. Unsupervised deep embedding for clustering analysis. In: Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR, 2016. 478−487

109 Langner O, Dotsch R, Bijlstra G, Wigboldus D H J, Hawk S T, Van Knippenberg A. Presentation and validation of the Radboud Faces Database. *Cognition and Emotion*, 2010, **24**(8): 1377−1388

110 Guo Y D, Zhang L, Hu Y X, He X D, Gao J F. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 87−102

111 Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD birds-200-2011 dataset [Online], available: http://www.vision.caltech.edu/visipedia/CUB-200-2011.html, November 6, 2011

112 Zhu Y, Tian Y D, Metaxas D, Dollár P. Semantic amodal segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 3001−3009

113 Borji A, Izadi S, Itti L. iLab-20M: A large-scale controlled object dataset to investigate deep learning. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 2221−2230

114 Marti U V, Bunke H. The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 2002, **5**(1): 39−46

115 Liu H Y, Tian Y H, Wang Y W, Pang L, Huang T J. Deep relative distance learning: Tell the difference between similar vehicles. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 2167−2175

116 Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, et al. Sentinel-2: ESA′s optical high-resolution mission for GMES operational services. *Remote Sensing of Environ-*

ment, 2012, **120**: 25−36

117 LeCun Y, Huang F J, Bottou L. Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Washington, USA: IEEE, 2004. II−104

118 Charles J, Pfister T, Everingham M, Zisserman A. Automatic and efficient human pose estimation for sign language videos. *International Journal of Computer Vision*, 2014, **110**(1): 70−90

119 Shahroudy A, Liu J, Ng T T, Wang G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1010−1019

120 Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. Cambridge, UK: IEEE, 2004. 32−36

121 Liu Z W, Luo P, Qiu S, Wang X G, Tang X O. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1096−1104

122 Zhang W W, Sun J, Tang X O. Cat head detection - how to effectively exploit shape and texture features. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 802−816

123 Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **36**(7): 1325−1339

124 Zhang W Y, Zhu M L, Derpanis K G. From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 2248−2255

125 Krause J, Stark M, Deng J, Li F F. 3D object representations for fine-grained categorization. In: Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE, 2013. 554−561

126 Khosla A, Jayadevaprakash N, Yao B, Li F F. Novel dataset for fine-grained image categorization: Stanford dogs. In: Proceedings of the 1st Workshop on Fine-Grained Visual Categorization. Colorado Springs, USA: IEEE, 2011. 1−2

127 Reichert D P, Seriès P, Storkey A J. A hierarchical generative model of recurrent object-based attention in the visual cortex. In: Proceedings of the 21st International Conference on Artificial Neural Networks. Espoo, Finland: ICANN, 2011. 18−25

128 Johnson J, Hariharan B, Van Der Maaten L, Li F F, Zitnick C L, Girshick R. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1988−1997

129 Qi L, Jiang L, Liu S, Shen X Y, Jia J Y. Amodal instance segmentation with KINS dataset. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 3009−3018

130 Eastwood C, Williams C K I. A framework for the quantitative evaluation of disentangled representations. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018.

131 Wu Z Z, Lischinski D, Shechtman E. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 12858−12867

文载道 西北工业大学自动化学院副教授. 主要研究方向为压缩感知与稀疏模型, 认知机器学习, 合成孔径雷达图像解译, 多源自主目标识别.
E-mail: wenzaidao@nwpu.edu.cn
(**WEN Zai-Dao** Associate professor at the School of Automation, Northwestern Polytechnical University. His research interest covers compressed sensing and sparse model, cognitive machine learning, synthetic aperture radar image interpretation, and multisource automatic target recognition.)

王佳蕊 西北工业大学自动化学院博士研究生. 主要研究方向为解耦表征学习, SAR 图像处理, 因果推理.
E-mail: wangjiarui_wyy163@163.com
(**WANG Jia-Rui** Ph. D. candidate at the School of Automation, Northwestern Polytechnical University. Her research interest covers disentangled representation learning, SAR image processing and causal reasoning.)

王小旭 西北工业大学自动化学院教授. 主要研究方向为惯性器件与惯性导航, 合成孔径雷达图像解译, 协同感知. 本文通信作者.
E-mail: woyaofly1982@163.com
(**WANG Xiao-Xu** Professor at the School of Automation, Northwestern Polytechnical University. His research interest covers inertial devices and inertial navigation, synthetic aperture radar image interpretation, cooperative sensing. Corresponding author of this paper.)

潘 泉 西北工业大学自动化学院教授. 主要研究方向为信息融合理论及应用, 目标跟踪与识别技术, 光谱成像及图像处理.
E-mail: quanpan@nwpu.edu.cn
(**PAN Quan** Professor at the School of Automation, Northwestern Polytechnical University. His research interest covers information fusion theory and application, target tracking and recognition technology, spectral imaging and image processing.)