Journal of Computer Applications
计算机应用, 2021, 41(12): 3409 - 3418
ISSN 1001-9081
CODEN JYIIDU
2021-12-10
http://www.joca.cn

文章编号:1001-9081(2021)12-3409-10
DOI:10.11772/j.issn.1001-9081.2021060895

解耦表征学习研究进展

成科扬[1,2*],孟春运[1],王文杉[2],师文喜[2,3],詹永照[1]
(1. 江苏大学 计算机科学与通信工程学院,江苏 镇江 212013;
2. 社会安全风险感知与防控大数据应用国家工程实验室(中国电子科学研究院),北京 100041;
3. 新疆联海创智信息科技有限公司,乌鲁木齐 830011)
(*通信作者电子邮箱kycheng@ujs.edu.cn)

摘　要:解耦表征学习旨在对影响数据形态的关键因素进行建模,使得某一关键因素的变化仅仅引起数据在某项特征上的变化,而其他的特征不受影响,这有利于应对机器学习在模型可解释性、对象生成和操作以及零样本学习等问题上的挑战,因此解耦表征学习一直是机器学习领域的一个研究热点。从解耦表征学习的历史与动机入手,对解耦表征学习的研究现状以及应用进行归纳总结,分析了解耦表征所具有的不变性、复用性等特性,介绍了基于生成解耦表征变差因素的研究、基于流形相互作用解耦表征变差因素的研究、基于对抗性训练解耦表征变差因素的研究,以及一种变分自编码器β-VAE的研究等最新研究动态。同时,阐述了解耦表征学习的典型应用,并对未来的研究方向作出了展望。

关键词:解耦学习;表征学习;变分推断;可解释性;机器学习;自编码器;变差因素;复用性
中图分类号:TP391　　**文献标志码**:A

Research advances in disentangled representation learning

CHENG Keyang[1,2*], MENG Chunyun[1], WANG Wenshan[2], SHI Wenxi[2,3], ZHAN Yongzhao[1]
(1. *School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*;
2. *National Engineering Laboratory of Big Data Application for Social Security Risk Perception and Prevention by Big Data (China Academy of Electronic and Information Technology), Beijing* 100041, *China*;
3. *Xinjiang Lianhai Chuangzhi Information Technology Company Limited, Urumqi Xinjiang* 830011, *China*)

Abstract: The purpose of disentangled representation learning is to model the key factors that affect the form of data, so that the change of a key factor only causes the change of data on a certain feature, while the other features are not affected. It is conducive to face the challenge of machine learning in model interpretability, object generation and operation, zero-shot learning and other issues. Therefore, disentangled representation learning always be a research hotspot in the field of machine learning. Starting from the history and motives of disentangled representation learning, the research status and applications of disentangled representation learning were summarized, the invariance, reusability and other characteristics of disentangled representation learning were analyzed, and the research on the factors of variation via generative entangling, the research on the factors of variation with manifold interaction, and the research on the factors of variation using adversarial training were introduced, as well as the latest research trends such as a Variational Auto-Encoder (VAE) named β-VAE were introduced. At the same time, the typical applications of disentangled representation learning were shown, and the future research directions were prospected.

Key words: disentangled learning; representation learning; variational inference; interpretability; machine learning; auto-encoder; factors of variation; reusability

0　引言

表征是机器学习中最基本的问题之一,它涉及视觉、语音识别、自然语言处理、强化学习和图形学等多个领域。然而,什么才是一个好的表征是一个存在争议性的问题。目前,表征有两种形式:一种是设计出来的,另一种是从数据中学习的。设计的表征可以完美地满足结构化复用和可解释性的需求;而学习的表征不需要专业知识,几乎在每一个具有足够数据的任务上都优于设计表征的功能,并且学习表征的特点取决于它们的预期用途。

近年来,机器学习取得了显著的进展,特别是在监督学习和强化学习方面。然而,许多性能优秀的算法常常存在数据

收稿日期:2021-05-12;**修回日期**:2021-06-21;**录用日期**:2021-06-25。
基金项目:国家自然科学基金资助项目(61972183,61602215);社会安全风险感知与防控大数据应用国家工程实验室主任基金资助项目。
作者简介:成科扬(1982—),男,江苏南通人,教授,博士,CCF会员,主要研究方向:计算机视觉、模式识别;　孟春运(1994—),男,江苏扬州人,硕士研究生,主要研究方向:计算机视觉、模式识别;　王文杉(1994—),女,湖北武汉人,硕士,主要研究方向:统计分析、机器学习;　师文喜(1988—),男,北京人,博士,主要研究方向:大数据分析、智慧安防;　詹永照(1962—),男,福建尤溪人,教授,博士生导师,博士,主要研究方向:多媒体、人工智能。

效率低下的问题，这些算法的表现往往缺乏生物智能所特有的鲁棒性和泛化性。为解决单个任务的上述缺陷，应该着力于解决主要特征对该任务适用性的问题。从实践角度来看，为每一项这样的任务去学习单独的表征需要重复设计模型而且浪费大量的时间；从理论角度来看，具有可分解结构、与不同部分相关联的一致语义的表征更有可能推广到新任务。在深度学习领域，已经进行了相当一段时间的研究来学习数据中的变差因素，通常被称为解耦表征学习。这一术语虽然没有规范的定义，但通常采用2013年Bengio等[1]的定义：一种表征，其中一个维度的变化对应于一个变化因子的变化，而其他因子相对不变。通俗来讲，解耦表征学习是对影响数据形态的关键因素进行建模，使得某一关键因素的变化仅仅引起数据在某项特征上的变化，而其他的特征不受影响，在模型可解释性、对象生成[2]和操作以及零样本学习等问题上有着巨大优势。

从用户的角度来看，解耦表征学习在图像处理方面具有强大优势。相较于传统的美图技术，解耦表征在此方面的功能更加丰富，生成的照片比基于生成对抗模型的应用更加逼真。从开发人员的角度来看，解耦表征在可解释性方面存在显著优势，使得基于此技术的应用场景更加广阔。如在医疗、金融领域，深度学习因不具备可解释性的缘故，一旦结果出现误差（医疗事故），无法解释误差的来源，以及金融部门对不透明模型的广泛应用可能导致的缺乏解释性和可审计性表示担忧。

由此可见，解耦表征学习研究意义重大，使深度学习这样的黑盒实验变得具有可解释性，研究者可以清楚地知道每一层特征提取的具体含义以及每一个决策背后的逻辑推理。本文将从解耦表征学习的历史与优势着手，并对解耦表征学习的研究现状以及应用进行归纳总结，同时对解耦表征学习的前景作出展望。

1 解耦表征学习研究的历史与优势

1.1 解耦表征研究探索期

1.1.1 解耦表征研究的起源

随着心理学和人工智能领域的蓬勃发展，研究人员试图将这样一种观点正式化，即世界上不同的变差因素可以从原始的感知输入中恢复。在早期视觉研究中认为某些变换会使对象特性不变，且这样的变换都是相互独立的，并随着这些变换的进行，视觉系统如何形成不变的表征对于理解感知至关重要。例如1979年J. J. Gibson所写的《视觉感知的生态方法》[3]一书中有关于形成不变表征方面的介绍。继Gibson的研究之后，心理学家立刻就注意到了群论的数学框架，将其作为一种把不变量和对称性概念正式化的方法[4]。

许多关于感知和表征学习的研究，特别是在物体识别方面，都遵循Gibson等[3]所强调的姿势或光照等特征，这些特征变换具有独立性[5-8]，在此框架中，变换被视为要丢弃的变量。然而，其他研究人员提倡一种表征学习的方法来保留这些变换的信息[9]。在感知的等价变换方法中，特征的一些子集对于某些特定的变换可能是不变的，但总体表征仍保留所有信息。特别是在无监督的环境中，这些表征可能用于许多不同的任务，因此等价变换的感知方法似乎更有效。解耦表征学习的工作就属于这一研究范畴，标志着解耦表征学习的起源。

1.1.2 不变特征对解耦表征的推动

在许多机器学习任务中，数据来源于涉及多因素复杂交互的生成过程。单独来看，每个因素都是数据可变性的一个来源。各个因素的交互共同带来了丰富的结构特征。为了应对这些变化因素，在机器学习和诸如计算机视觉的应用领域中出现了一种研究热潮，即转向研究数据变化源中保持不变的手工工程特性集。这股热潮推动了LeCun等[10]在1989年提出卷积网络结构中特征池化阶段的表征和2009年Wang等[11]基于大规模低级别特征池化表征的发展。这些方法都源于一个想法，即通过汇集一组简单的滤波响应来归纳数据的不变特征。2011年Courville等[12]从无监督数据中弄清了哪些过滤器可以合并在一起，从而提取了汇集特征不变的方差方向。1996年Kohonen[13]提出了ASSOM（Adaptive-Subspace Self-Organizing Map）模型，首次确立了这样的原则，即不变特征实际上可以通过无监督学习的方式从特征的子空间中产生。从那时起，相同的基本策略出现在一些不同的模型和学习范式中，包括2000年Hyvärinen等[14]提出的拓扑独立成分分析、2009年Kavukcuoglu等[15]提出的不变预测稀疏分解以及2010年Ranzato等[16]提出的基于玻尔兹曼机的方法。在每种情况下，基本策略都是通过使用一个变量来将过滤器分组在一起，该变量为组中所有元素的激活提供门控。对于不变特征的研究推动了解耦表征的发展，并为接下来变分自编码器（Variational Auto-Encoder, VAE）在解耦表征中的应用提供了良好的理论基础。

1.2 解耦表征对机器学习的影响

解耦表征学习实际上是将传统数学建模和机器学习建模的优缺点相结合的方法。传统数学建模方法是数学家们基于专家经验和对现实世界的理解进行建模，而机器学习建模则是另一种完全不同的建模方式。机器学习算法以一种更加隐蔽的方式来描述一些客观事实，尽管人类并不能完全理解模型的描述过程，但在大多数情况下，机器学习模型要比人类专家构建的数学模型更加精确。更重要的是，在诸多应用领域（如医疗、金融、军事等）[17]，机器学习算法尤其是深度学习模型并不能满足当下需要清晰且易于理解的决策需求。

深度学习技术的发展在许多领域都对神经网络的应用进行了尝试。在一些重要的领域，使用神经网络确实是合理的，并且获得了较好的应用效果，包括计算机视觉、自然语言处理、语音分析和信号处理等[18-21]。在上述应用中，深度学习方法都是利用线性和非线性转换对复杂的数据进行自动特征提取，并将特征表示为“向量”（vector），这一过程一般也称为“嵌入”（embedding）。之后，神经网络对这些向量进行运算，并完成相应的分类或回归任务。从特征提取和准确度来看，这种“嵌入”的方法非常有效，但在许多方面也存在不足，如：

1）可解释性：嵌入所使用的N维向量无法对模型分类的原理和过程进行很好的解释，只有通过逆向工程才能找到输入数据中对模型分类影响较大的内容。

2）数据需求量庞大：如果只有10~100个样本，深度学习无法使用[22-23]。

3）无监督学习：大多数深度学习模型都需要有标签的训练数据。

4）零样本学习：这是一个很关键的问题，基于一个数据集所训练出的神经网络，若不经过重新训练，很难直接应用在另一个数据集上。

5)对象生成:除了生成对抗网络(Generative Adversarial Network, GAN)以外,其他模型都很难生成一个真实的对象[24-27]。

6)对象操作:难以通过嵌入调整输入对象的具体属性。

7)理论基础:虽然目前已经掌握了比较通用的逼近理论,但还远远不够。

针对这些棘手的问题,数学建模恰好可以解决。在几十年以前,大多数数学家都没有遇到过上述问题,因为他们主要关注数学建模(mathematical modeling),并通过数学抽象来描述现实世界中的对象和过程,如使用分布、公式和各种各样的方程式。在这个过程中,数学家定义了常微分方程(Ordinary Differential Equation, ODE)。本文通过指出深度学习存在的问题,对数学建模的特点进行了分析。需要注意的是,"嵌入"代表数学模型的参数,如微分方程的自由度集合。对于机器学习的不足,相应的数学建模在上述几个方面都得到了有效的解决,如:

1)可解释性:每个数学模型都是基于科学家对客观事物的描述而建立的,建模过程包含数学家对客观事物的描述动机和深入理解。

2)数据需求量:大多数数学建模上的突破并不需要基于巨大的数据集进行。

3)无监督学习:对数学建模来说也不适用。

4)零样本学习:一些随机微分方程(如几何布朗运动)适用于金融、生物或物理领域,只需要对参数进行重新命名。

5)对象生成:不受限制,对参数进行采样即可。

6)对象操作:不受限制,对参数进行操作即可。

7)理论基础:百年数学的奠基。

如果在处理复杂数据时把表现较好的神经网络和人工建模方法结合起来,可解释性、生成和操作对象的能力、无监督特征学习和零样本学习的问题,都可以在一定程度上得到解决。对于微分方程和其他人工建模方法来说,图像处理很难进行,但通过和深度学习进行结合,上述模型允许我们进行对象的生成和操作、增强模型的可解释性[28-30],最重要的是,该模型可以在其他数据集上完成相同的工作。但该模型也有一些不足之处,如建模过程不是完全无监督的,目前半监督、自监督的方法很好地弥补了这一短板[31-33]。

下面本文将着重介绍分析VAE这个结合了传统数学建模和机器学习优缺点的模型来实现解耦表征,也就是让嵌入中的每个元素对应一个单独的影响因素,并能够将该嵌入用于分类、生成和零样本学习。

2　解耦表征学习的研究方向与现状分析

2.1　变差因素方面的研究现状

解耦表征学习不同于分类等其他机器学习任务,其难点之一是难以建立明确的目标或训练目标。在分类任务的情况下,目标是显而易见的,即把训练数据集上的错误分类数量降至最低。在解耦表征学习的情况下,其目标与最终目标相去甚远,最终目标通常是学习分类器或其他预测器,而解耦表征学习的目标是一种解开变化的潜在因素的表征;同时另一个值得考虑的问题是如何将其转化为合适的训练标准。下面从变差因素这个研究方向进行归纳总结:一方面,这属于解耦表征学习的主要研究方向,从中可以了解到目前的研究现状;另一方面,也是解耦表征学习的主要应用,在有限的数据资源中减少模型的训练次数并获得更大的泛化效果。

2.1.1　基于生成耦合解开变差因素的研究

2011年Courville等[34]提出了钉板约束的玻尔兹曼机(spike-and-slab Restricted Boltzmann Machine, ssRBM)被证明是一种很有前途的自然图像数据模型。在这里,为了学习如何理清数据中明显的变差因素,Courville等将ssRBM推广到多个二元隐变量的高阶相互作用中。从生成角度来看,二元隐变量的乘法相互作用模拟了产生数据因素的耦合。相反,模型中的推理可以被视为试图将各种相互作用的因素归因于它们对数据的综合描述,从而在效果上解耦变差因素。这种方法只依赖模型参数的无监督近似最大似然学习,因此在定义要解耦的因素时不需要使用任何标签信息。

基于生成耦合解开变差因素的方法源于钉板约束的玻尔兹曼机,并将其推广到多个潜在变量的高阶相互作用中[35]。从生成的角度看,乘法的相互作用模拟了变差因素的耦合。与以往的潜在因素分析方法不同,基于生成耦合解开变差因素的方法不使用潜在因素的监督信息进行训练,例如考虑面部表情识别的目的。一方面,具有相同面部表情的不同个体的两个图像可能会使图像在像素空间中被很好地分开;另一方面,显示不同表情的相同个体的两个图像很可能在像素空间中被放置在非常接近的位置。重要的是,这些相互作用的因素通常不会组合成简单的叠加,通过选择适当的数据仿射投影就可以很容易地将其分开,这些因素往往看起来与原始数据紧密相关,这其实是一种基于生成耦合解开变差因素的研究方法。

基于生成耦合解开变差因素的进一步研究是2011年Hinton等[36]提出通过学习提取与姿势参数相关联的特征来分离变差因素,其中姿势参数的变化在训练时是已知的。该模型也与2010年Memisevic等[37]提出的高阶玻尔兹曼机用作图像中的空间变换模型密切相关。虽然这两种模型有许多不同之处,最显著的区别是前者使用了潜在变量的乘法交互作用。虽然在玻尔兹曼能量函数中包含了更高阶的相互作用,但只在观察到的变量之间使用,极大地简化了推理和学习过程。如上所述,使用基于生成耦合解开变差因素的方法可以分离出相互作用的潜在变量组。以这种方式保持交互的局部性,是仅使用无监督数据成功学习的关键组成部分。

2.1.2　基于流形相互作用解耦变差因素的研究

2000年Tenenbaum等[38]提出流形学习方法,通过学习低维结构或嵌入实现数据建模。现有的流形学习方法可以从单个人的人脸图像中学习本质上的低维结构,如视点流形,但要对复杂的高维流形建模,如数百万人的人脸图像空间,则变得具有挑战性。然而,正如2011年Rifai等[39]所建议的那样,深度学习已经证明在学习高维数据流形方面是有效的,但是对多个变差因素及其相互作用的流形进行联合建模仍然是一个挑战。有几个相关的工作使用了多个隐变量的高阶交互作用。例如,2000年Tenenbaum等[38]提出的双线性模型用来分离面部图像和语音信号中的风格和内容;2013年Tang等[40]提出的张量分析器,通过引入因子加载张量来对多组潜在因子单元的交互作用进行建模,从而扩展了因子分析,并被应用于光照和脸部形态的建模。

一个能够具有解耦变差因素的生成模型应该能够保持遍历一个因素流形的同时固定其他因素的状态[41]。例如对于人脸图像生成模型,在固定身份的同时可以生成不同姿势或表

情的图片，并且还具备在子流形内插值，将固定的姿势或表情迁移给不同身份的人。2013 年 Bengio 等[42]指出，与像素或单层模型等浅层表征相比，跨越深层表征的线性插值可以更接近图像流形。当输入图像是由多个变差因素生成时，它们往往位于一个复杂的流形上，这使得学习有用的表征具有较高的难度。通过将每个变差因素视为自身形成一个子流形来处理这个问题，并对各因素之间的联合作用进行建模。如图 1 所示，从不同方位角（固定高度）拍摄的单个人的人脸图像时，图像的轨迹将形成环形。

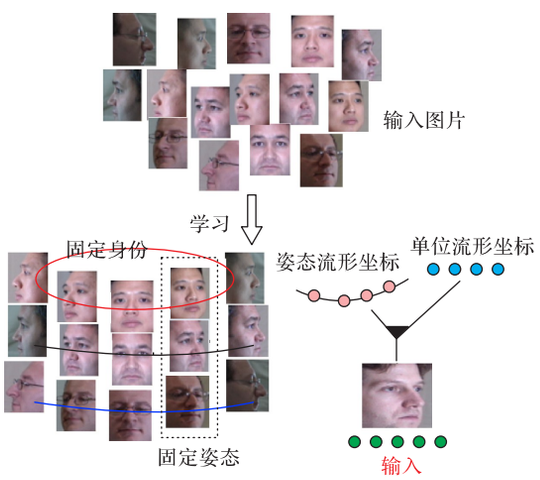


图 1 面部图像中建模姿势和身份变化的方法

Fig. 1 Approach for modeling pose and identity variations in face images

许多潜在的变差因素相互作用可产生视觉数据，因此，可以学习相关变差因素的流形坐标，并对它们的联合作用进行建模。现有的许多特征学习算法都集中在单个任务上，提取对任务的相关因素敏感、对其他因素不变的特征。然而，只提取一组不变特征的模型并没有很好地利用潜在因素之间的关系。为了解决这个问题，2014 年 Reed 等[43]提出了一种高阶玻尔兹曼机，它结合了隐藏单元组之间的乘法交互作用，每个隐藏单元组都学习编码不同的变差因素。此外，Reed 等还提出了基于通信的训练策略模型，该模型在多伦多人脸数据库上实现了高效的情感识别和人脸验证性能。

2.1.3 运用对抗性训练解耦深层表征变差因素的研究

2009 年 Bengio 等[44]认为理解感官数据的一个根本挑战是学会解开潜在的变差因素。例如，生成语音记录所涉及的变差因素包括说话人的属性，如性别、年龄或口音，以及语调和正在说话的内容。类似地，图像的潜在变差因素包括图像的物理表现和观察条件。解开这些隐藏因素的困难在于多数情况下，每一个因素都会以不同的、不可预测的方式影响观察，很难有人可以将接触到这些影响的性质明确给出标签数据。通常情况下，收集数据集的目的是进一步解决某个有监督的学习任务，且这种类型的任务学习完全是由标签驱动的。对抗性训练解耦深层表征是让学习到的表征对手头没有提供信息任务的变差因素保持不变。2019 年 Li 等[45]提出了一种名为相似性限制的生成式对抗网络（Similarity Constraint Generative Adversarial Network，SCGAN），它能够以完全无人监督的方式学习解耦表征。他们的灵感来自平滑度假设以及对图像内容和表征的假设，并设计了有效的相似性约束。SCGAN 可以通过在条件和合成图像之间添加相似度约束来解耦出具有可解释性的表征。实际上，相似性约束是作为指导生成网络来理解基于条件表征差异的一种监督。SCGAN 成功区分了许多数据集的不同表征。虽然最近的监督学习方法取得了巨大的成功，但它们是以放弃解决其他相关任务的可复用性为代价的。

许多其他的应用需要使用生成模型，这些生成模型能够合成新的实例，并将某些关键的变差因素保持恒定。与分类不同，生成式建模需要保留所有变差因素。但对于许多任务来说，仅仅保留这些因素是不够的。例如，在语音合成中将一个人的对话转换成另一个人的声音；以及图像处理中的逆问题，如去噪和超分辨率。基于上述考虑，2016 年 Mathieu 等[46]引入了一个条件生成模型来学习如何从一组标记的观测中解开隐藏的变差因素，并将它们分离成互补码。一份代码总结了与标签相关的特定变差因素，另一份总结了剩余的未指明的可变性。在训练期间，唯一可用的监督来自区分同一类别的不同观察能力。这种观察力包括在不同视角捕获的一组标记对象的图像，或者口述多个短语的一组说话者的记录。在这两种情况下，类内差异是未指明的变差因素的来源：每个对象在多个视角被观察，每个说话者口述多个短语。当进行强有力的监督时，将特定因素与未指定因素区分开也变得更加简单快捷。然而，在数据采集不受严格控制的情况下，未指明因素的标签通常是不可用的，所以通过将深度卷积自编码器和另一种形式的对抗性训练相结合的方式来解决在更一般情况中的解耦问题。

2.2 变分自编码器变体方面的研究现状

目前，先进的解耦学习很大程度上基于 VAE：假定隐空间（latent space）特定的先验概率 $P(z)$，然后使用一个深度神经网络参数化条件概率 $P(x|z)$。类似的，分布 $P(x|z)$ 是使用变分分布 $Q(z|x)$ 近似得到，该变分分布也是由深度神经网络参数化得到的。通过最小化负对数相似度的近似来训练模型。$r(x)$ 的表示通常取近似后验分布 $Q(z|x)$ 的均值。为了获得更好的解耦性，几种不同的 VAE 模型被提出，但这些方法的共同点都是试图分解聚合后验 $\int_x Q(z|x)P(x)\mathrm{d}x$，这可以帮助解耦[47-48]。

生成模型能否学习到解耦表征的常见标准是能否将其数据集的形式和内容作为独立变差因素来进行学习。为了利用 VAE 实现这种解耦，标签信息通常以完全监督或半监督的方式提供，然而，变分目标不足以解释所观察到的形式及内容的解耦，因此，通过不同形式的自编码器的变种可以有效解决这种问题[49]。下面将介绍一些自编码器的变种在解耦表征方面的研究。

2.2.1 代表性的变分自编码器变体

在无监督的情况下学习解耦表征是有难度的。分离的变量通常被认为包含可解释的语义信息和独立作用于生成过程的组合。与主成分分析（Principal Component Analysis，PCA）类似，一方面自编码器最初被视为一种降维技术，因而在使用上有一定瓶颈；另一方面，稀疏编码和限制玻尔兹曼机（Restricted Boltzmann Machine，RBM）方法的成功使用倾向于支持完整表征，这可以允许自编码器简单地复制特征中的输入，具有完美的重构。在此基础上，2013 年，Kingma 等[50]提出了 VAE，该模型还能够捕捉到图像的结构变化（倾斜角度、圈的位置、形状变化、表情变化等）。2016 年 Higgins 等[51]提出的

β-VAE是一种无监督视觉解耦表征学习模型,它是对VAE目标函数的修改,是为了了解图像x及其潜在生成因子z联合分布的一种生成方法。β-VAE在VAE目标上增加了一个额外的超参数β,这限制了潜在瓶颈的有效编码能力,并鼓励了潜在表征的更多因式分解。β-VAE所学习的解耦表征对于学习抽象视觉概念层次[52]、提高强化学习策略的迁移性能(包括机器人学中模拟现实迁移)具有重要意义。β-VAE有助于将解耦因子学习扩展到更复杂的数据集上。特别是与标准的VAE相比,β-VAE学习的因子表征可以与人类对数据生成因子的直觉轴对齐。不过,β-VAE也有其局限性,如重建保真度比标准VAE差。这是由修改后的训练目标引入的权重导致的,该目标函数变相地降低了重建损失的权重,以便鼓励在潜在表征中解耦。与标准的VAE目标相比,Higgins等[53]对β-VAE学习视觉数据生成因素的轴对齐解耦表征的原因有了新的见解。特别是鼓励β-VAE找到一组代表性轴,这些轴能够更好地保持数据点的局部性,并且与提高数据记录可能性的变差因素保持一致。该方法通过允许先验的平均KL(Kullback-Leibler)散度从零逐渐增加,而不是原始β-VAE目标中固定的β加权KL项,因此使得训练过程中可以控制潜在后验编码容量的增加。结果表明,与原始公式中获得的结果相比,这提高了解耦表征的鲁棒性和更好的重建保真度。2018年Burgess等[54]对VAE中的解耦表征提出了新的理论评估方法,从速率失真理论的角度展示了当修改证据下界优化β-VAE时,随着训练的进行,出现了与数据变化的潜在生成因素相一致的表征的情况。基于这些认识,对β-VAE的训练机制提出了一种改进方案,即在训练过程中逐步增加潜码的信息容量。这种修改加强了β-VAE中解耦表征的鲁棒性,而不需要过多考虑先前在重建精度方面的权衡。

学习数据语义是人工智能的一个重要分支。这种表征方式不仅对标准下行任务有用,如监督学习和强化学习,还对迁移学习和zero-shot学习这种人类胜于机器的学习有用。在深度学习领域,科研人员在数据的变差因素方面付出了艰苦努力。实际上,假设数据从一个有着固定数量的变差因素中生成,在集中处理图像数据的过程中,其变量中因素的影响更容易可视化。使用生成模型表明了在图像中学习解耦表征的巨大希望。2018年Kim等[55]定义并解决了对独立变差因素产生的数据进行解耦表征学习的问题,并提出了FactorVAE。算法通过激励表征的分布,使之成为因式,并在整个维度中独立。FactorVAE改进了β-VAE,在解耦和重构质量之间提供了更好的折中,并且引入了一种常用的解耦度量。图2是FactorVAE的架构,即一个可变的自动编码器,阶乘鼓励代码的分布(encouraging the code distribution to be factorial)。图2上方是一个带有卷积编码器和解码器VAE,下方是一个多层感知机(Multi-Layer Perceptron, MLP)分类器,即判别器,它是用来区分输入进来的是边缘码分布还是边长的乘积。

在VAE中,解耦就会导致重构不是很好,FactorVAE希望可以做到在不降低重构质量的情况下,得到较好的解耦效果。导致VAE解耦效果不佳主要归因于其KL散度分解完包含两个函数,即$I(x;\ z)$和$\mathrm{KL}\left(q(z)\,\|p(z)\right)$,前者互信息是指$z$中包含多少$x$的信息,后者是让$q(z)$与$p(z)$接近一点,由于$p(z)$各向同性的性质,$q(z)$向$p(z)$靠近会导致$z$的各维度之间也是独立的,这就是FactorVAE想要达到的目的;所以在β-VAE里面直接对两项一起惩罚会导致解耦效果变好,而重构效果就会下降,因此FactorVAE希望将其分开。最后,FactorVAE提出使用总相关惩罚系数(Total Correlation Penalty, TCP),就是在传统VAE的下界再加一TC项,它可以促进z的每个维度之间尽可能独立。

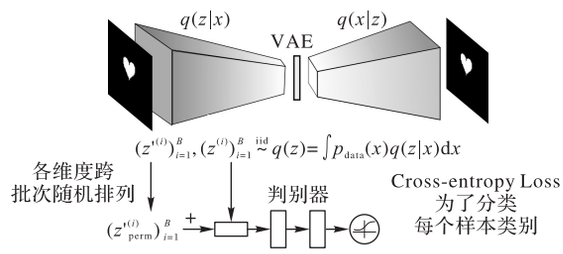


图2　FactorVAE的架构

Fig. 2　Architecture of FactorVAE

在过去的几年中,在利用生成模型解耦潜在表示方面取得了重要进展,其中两种主要方法是GAN和VAE。但值得注意的是,标准的自编码器和紧密相关的结构模型仍然很流行,因为它们易于训练并且可以适应不同的任务。所以引出了一个有趣的问题:能否让自编码器在保持其良好性能的同时实现优秀的解耦能力。2020年Patacchiola等[56]提出了新模型Y-AutoEncoder(Y-AE)来解决了上述的问题。Y-AE的结构将表征形式分为隐性和显性两个部分:隐性部分类似于AE的输出,显性部分与训练集中的标签密切相关。通过在解码和重新编码之前将编码器的输出分成两条路径(形成Y形),在潜在空间中将这两部分分开,然后施加多种损失,例如重建损失,以及隐性和显性部分之间依赖性的损失。从各个方面的实验结果来看,例如样式和内容的分离、图像到图像的转换等都有显著的提升。

2.2.2　不同解耦度量的侧重

不同的解耦模型针对不同的下游任务会导致各个模型解耦性能的差异。虽然对解耦表征希望有一个全面、统一的评估,但还没有一个单一、被普遍接受的定义。一般说来,如果不是完整的基本事实生成模型的话,所有这些方法都假设它们可以获得基本事实因素。在以下内容中,将简要回顾在研究中考虑的解耦度量。

2016年Higgins等[57]提出了β-VAE度量,建议在基本生成模型常数中固定随机变差因素,采样两个小批次的观测样本;然后,利用线性分类器的精度来表示模型的解耦性能,该线性分类器基于两个小批次中表征向量之间的平均差值来预测固定因子的指数。随后2018年Kim等[55]提出了FactorVAE度量,通过使用多数投票分类器来解决该度量的几个问题,该分类器基于具有最小方差的表征向量的索引来预测固定基础事实因子的索引。同年Chen等[58]提出了互信息间隔(Mutual Information Gap, MIG),即具有最高互信息的表征因子和次高互信息的表征因子之间的归一化均方误差,并认为β-VAE度量和FactorVAE度量并不具备一般性,因为它们依赖于一些超参数。Chen等比较表征中的每个基本事实因子和其他因子的成对互信息,然后考虑其中具有最高互信息的两个因子。2018年Ridgeway等[59]提出了应该考虑表征的两个不同性质,即模块性和显性。在模表征中,表征出来的每个维度至多取决于单个变差因素;在显式表征中,变差因数的值很容易从模表征中预测。Ridgeway等提出用模表征的互信息间隔来衡量模数。在测量显性时,利用一对逻辑回归分类器的ROC-AUC(Receiver Operating Characteristic-Area Under Curve)来预测变

差因素。2019 年 Kumar 等[60]提出了分离属性可预测性(Separated Attribute Predictability, SAP)分数,即从学习表征的每个维度预测因子值中计算线性回归的 R2 分数(决定系数)。对于离散数据的情况,他们建议训练分类器。SAP 分数是每个因素的两个最具预测性的潜在维度的预测误差的平均值。

通常情况下,比较模型性能的优劣应当使用普遍接受的度量方法。目前,解耦表征针对不同研究方向提出了各种不同的具体度量标准,每个度量标准的目的都是以稍微不同的方式来形式化度量的概念。

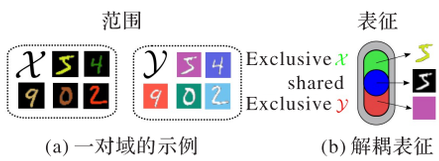
3 解耦表征学习研究的应用与展望

3.1 解耦表征的实际应用

3.1.1 解耦表征从图像到图像转换上的应用

深度学习极大地提高了图像到图像转换方法的质量。这些方法旨在学习将图像从一个域转换到另一个域的映射。如:彩色化,其目的是将灰度图像映射到相同场景的彩色图像[61];语义分割,其目的是将 RGB 图像被转换为指示 RGB 图像中的每个像素的语义类别的映射[62]。但是这类映射有两个主要挑战:一是缺少对齐的训练对;二是来自单个输入图像的多种可能的输出。2017 年 Isola 等[63]提出了一种从图像到图像转换的通用方法。当配对数据可用时,该方法可以成功应用于各种问题。通过引入循环一致性损失,该理论进一步扩展到未配对数据的情况下[64]。U-Net 架构[65]通常用于图像到图像的转换。该网络可以解释为编码解码器网络,编码器从输入域提取相关信息,并将其传递给解码器,然后解码器将该信息转换到输出域。尽管这些模型目前很流行,但学习到的表征(编码器的输出)只针对一些特定任务[66]。然而,解耦表征可以对图像到图像转换模型中学习到的特定表征进行结构化。将偶然的场景事件(如光照、阴影、视点和对象方向)从场景的固有属性中分离出来一直是计算机视觉希望达成的目标[67]。当应用于深度学习时,会使得深度模型能够意识到独立的变化因素[68]影响所表示的实体。因此,如果信息与手头的任务不相关,模型可以沿着特定的变化因素将信息边缘化。这样的过程对于被特定因素(例如,对象识别中变化的照明条件)而阻碍的任务将会大有益处。

如图 3 所示,左边为一对域的示例,其中包含黑色背景上带有彩色数字或彩色背景上带有白色数字的图像;右边为解耦表征,分为跨域共享部分(数字)和域独占部分(背景或数字上的颜色)。将解耦目标与图像到图像的转换相结合,引入了跨域解耦表征的概念,其目的是将特定领域的因素从跨域共享的因素中分离出来。为此,将表征划分为三个部分:共享部分包含两个域共有的信息,而两个独占部分仅表示每个域特有的变化因素。2018 年 Gonzalez-Garcia 等[69]提出了基于跨域的双向图像转换模型,该模型使用一对生成性对抗网络,通过充分结合多个损失和一个称为跨域自编码器的新网络组件,在学习表征中实施解耦结构。由于解耦表征具有更好的可控性和通用性,因此取得了比现有方法[70]更好的结果。为解决图像转化过程中缺少对齐的训练对且出现多种可能的输出问题,2020 年 Lee 等[71]基于解耦表征的方法,在未配对训练图像的情况下综合了各种输出。该模型采用从给定输入中提取的编码内容特征和从属性空间中采样的属性向量,以在训练时综合各种输出;并且引入了一种跨周期一致性损失函数来处理未配对的训练数据。结果表明,该模型可以在没有配对训练数据的情况下在广泛的任务中产生多样化的现实图像。用于图像到图像转换的跨域解耦表征模型[72]具有几个优点:1)样本多样性,可以根据输入图像生成图像的分布,而大多数图像到图像架构只能生成确定性结果,虽然明确地对两个域中的变化进行建模,而它们只考虑输出域中的变化;2)跨域检索,可以基于域之间共享表征的部分来检索两个域中的相似图像,而不需要标记数据来学习共享表征;3)特定域的图像传输,特定域的特征可以在图像之间传输;4)特定域的插值,可以基于特定域在两个图像的特征之间进行插值。



(a) 一对域的示例 (b) 解耦表征

图 3 图像到图像的转换

Fig. 3 Image-to-image translation

3.1.2 解耦表征在视频中进行无监督学习的应用

从视频中进行无监督学习是计算机视觉和机器学习的一个研究热点。其目标是在不使用显式标签的情况下学习一种表征方法,这种方法可以有效地概括以前未见过的任务范围,例如对当前对象的语义分类、预测视频的未来帧数或对正在发生的动态活动进行分类。有几种流行的范例:第一种称为自我监督,使用领域知识隐式地提供标签(例如,预测对象上补丁的相对位置或使用特征轨迹),这允许将问题作为自生成标签的分类任务来处理;第二种方法依赖于辅助动作标签,在真实或模拟的机器人环境中可用,这种方法可以用来训练未来帧的动作条件预测模型[73],也可以用来训练从当前和未来帧中对预测动作的反向运动学模型[74];第三种也是最普遍的方法,即预测性的自编码器,试图根据当前帧预测未来帧。为了学习有效的表征,需要对潜在表征进行某种约束。

这里引入一种预测性的自编码器[75],它使用一种新的对抗性损失来将每个视频帧的潜在表征分解为两个分量:一个分量大致与时间无关(在整个剪辑中大致不变);另一个分量捕捉序列的动态方面,因此能够实现随时间而变化。这里分别将它们称为内容组件和姿态组件。对抗性的损失依赖于这样一种直觉,即尽管内容特征应该针对于特定剪辑,但个体的姿态特征不应该如此。因此,损失函数鼓励姿势特征携带没有关于剪辑身份的信息。从经验上看,这种损失的训练对于诱导所需的因式分解是至关重要的。两个独立的编码器为每个帧产生内容和姿势的不同特征进行表征,通过帧 x^t 的内容表征和未来帧 x^{t+k} 的姿势表征组合并解码以预测未来帧 x^{t+k} 的像素来训练它们。然而,仅此重构约束不足以在两个编码器之间诱导所需的因子分解。因此,2019 年徐思[76]在姿势特征上引入了一种新的对抗性损失,从而确保了它们不能包含内容信息,且另一个约束是内容信息应该随时间缓慢变化,这一概念鼓励在时间上相近的内容向量彼此相似。利用视频的时间相干性和一种新的对抗性损失来学习一种表征方法,该方法将每一帧分解为一个固定部分和一个时间变化的组件。因此,与该方法类似的解耦表征可用于一系列任务,如将标准长短期记忆(Long Short-Term Memory, LSTM)应用于时变分量可实现对未来帧的预测。但是这种方法无法保证网络学习到低耦合度的特征,所以 2020 年 Sun 等[77]提出了循环表征分

解网络来解决这一问题,采用属性交换和视频重构的策略将视频分解成静止和时间变化的分量,并在几个视频理解任务上获得了比现有技术更好的结果,在技术上做到了从视频理解的角度进行视频解耦表征学习。

3.1.3　解耦表征在3D建模方面的应用

深度生成模型可以合成逼真的图像,然而,大多数方法只专注于生成2D图像,而忽略了世界的3D本质。因此,用于生成2D图像的模型无法回答一些对人类来说毫不费力的问题,例如:从不同的角度看汽车会是什么样子?或者如果把小汽车的纹理应用到一辆卡车上呢?又或者可以混合不同的3D设计吗?因此,2D视角不可避免地限制了模型在机器人、虚拟现实和游戏等领域的实际应用。

2018年Zhu等[78]提出了一种端到端的生成模型——可视对象网络(Visual Object Network, VON),该模型能够将物体的自然图像与解耦3D表征的形式结合起来,联合生成3D形状和2D图像。具体地说,借鉴经典图形渲染引擎的思想,将图像生成模型分解为三个条件独立的因素:形状、视角和纹理。该模型首先学习生成像真实物体形状一样几乎难以区分的3D形状;然后用一个可微投影模块从一个采样的视角计算它的2.5D草图;最后,在2.5D草图上添加不同的、逼真的纹理,并生成与真实照片难以区分的2D图像。

同时,条件独立连接减少了密集标注数据的需求,与经典的可变形面部模型不同,可视对象网络的训练不需要2D图像和3D形状之间的配对数据,也不需要3D数据中密集的对应标注。这一优势使可视对象网络能够利用2D图像数据集和3D形状集合,并生成不同形状和纹理的对象。

3.2　解耦表征未来的研究方向

3.2.1　对归纳偏好和隐性、显性监督的探索

无监督解耦表征学习背后的想法:真实世界的数据是由一些可解释的变差因素所生成的,并且这些变差因素可以通过无监督学习算法来恢复。因此,使得模型更好的解耦是各类解耦方法的主要动机。这就提出了一个问题,即"模型选择、超参数选择和随机性(以不同随机种子的形式)是如何影响解耦效果的?"为了研究这个问题,2018年Locatello等[79]为训练过的模型计算出分离式度量,定量分析并归纳偏好设置。他们展示了Cars3D数据集上每种方法的FactorVAE得分范围,并且这些范围对于不同的模型是严重重叠的,所以得出这样的结论:超参数和随机种子的选择似乎比目标函数的选择重要得多。他们还进一步展示了随机种子形式的随机性对解耦效果的影响,并表示随机性(以不同的随机种子的形式)对获得的结果有很大的影响。

由于目前在解耦表征的无监督学习中似乎没有一个可靠的选择超参数的策略,纯粹的无监督解耦学习似乎从根本上是不可能的。2020年Locatello等[80]首先从理论上证明了如果没有模型和数据的归纳偏好,无监督的解耦表征学习是不可能的。从他们的实验结果来看,尽管使用不同的方法都成功地实现了由相应的损失函数优化出对应的属性,但似乎没有监督就无法确定很好的解耦模型。此外,不同的评估指标并不总是在解耦的问题上达成一致,且在估算中表现出系统性差异。最后,模型解耦性的提高似乎并没有必然导致下游任务的学习样本复杂度降低。所以有关解耦表征学习的未来工作应明确归纳性偏好和(隐性、显性)监督的作用,加强对所学解耦表征具体好处的探索,并考虑涵盖多个数据集可重现的实验设置。目前,具有交互性的解耦学习是一种很有前景的学习方式,如利用弱监督或数据时序结构用于学习问题。考虑到非线性独立成分分析(Independent Component Analysis, ICA)模型中的可识别性结果,对数据时序结构进行解耦表征学习似乎特别有趣,这可能表明,如果能够利用数据的顺序结构,则基于自动编码的方法可能会有显著的改进。

3.2.2　对解耦表征可解释性、公平性、互动环境及可复现实验设置方面的探索

未来的工作应该着眼于放大解耦表征学习的几个具体优势,如可解释性、公平性以及互动环境等方面,包含归纳偏好、提供可解释性和概括性的一个潜在方法是开创独立因果机制和因果推理的框架。2019年Van Steenkiste等[81]进行了一个大规模的实验研究,以实验解耦表征是否更适合于抽象推理任务。因此要想强调对各种数据集进行合理、可靠、可重复的实验设置,以便得出有效的结论,首先,必须谨慎对待实验设计,为所有不同的方法和解耦度指标选择了完全相同的训练和评估协议。其次,如果仅考虑方法、指标和数据集的子集,很容易从实验结果中得出虚假结论,数据集的更改可能会导致得出截然不同的结论,因此,对未来的工作而言,对各种数据集进行实验以查看结论和见解是否普遍适用至关重要。这在解耦学习的设置中特别重要,因为实验主要是在类似玩具数据集上进行的,因此,目标应当是对多个数据集进行概括的能力,而不是对特定数据集的绝对性能。

3.2.3　对解耦表征深层理论和评价指标方面的探索

解耦表征学习的关键是从观察中理解世界、在不同任务和领域之间转移知识和学习组成概念。虽然近年来取得了相当大的进展,但关于陈述解耦表征的常见假设的合理性似乎是不够的。至少有两个主要原因导致目前的解耦表征状态不令人满意:一是缺乏正式的解耦表征概念来支持适当目标函数的设计;二是缺乏有效的评估度量来实现模型之间的公平比较。

2020年Do等[82]从信息性、可分离性和可解释性三个维度描述了在监督和无监督方法中使用解耦表征的概念,这些概念可以用信息论的结构明确地表示和量化。然后,他们还提出了用于度量信息性、可分离性和可解释性的评价指标。目前,在解耦表征的理论和评价指标这两方面的探索还处在初步阶段,鼓励设计更多理论驱动的模型来探索解耦表征是未来非常重要的一个研究方向。

4　结语

人工智能技术的蓬勃发展,改变了人们生活的方方面面。机器和人类在很多复杂认知任务上的表现已经不分伯仲。然而,单一模型只能解决单一问题,目前学界对解耦学习的研究尚处于起步阶段。从当前研究现状看,研究者们普遍意识到提高模型利用率的重要性,并已展开了诸多有意义的研究。在未来,解耦表征学习的研究应该更加着力于对归纳偏好、无监督或者自监督学习的探索上来。对解耦表征可解释性、公平性、互动环境以及其深层理论和评估方式上还有很长的路要走。本文从解耦表征学习的历史与动机、研究方向和现状以及解耦表征学习的应用与展望这几方面对当前解耦表征学习的研究进行总结,让读者了解解耦表征学习的发展脉络、研究方向、取得的成果和未来的发展趋势。相信今后随着对解耦表征学习深入的研究,在模型的泛化能力、提升处理非单一

任务时的利用率以及深度学习可解释性等方面一定会有所突破。

参考文献(References)

[1] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[2] 胡铭菲,刘建伟,左信. 深度生成模型综述[J/OL]. 自动化学报. (2020-09-21)[2020-09-27]. https://doi.org/10.16383/j.aas.c190866. (HU M F,(LIU J W, ZUO X. Survey on deep generative models[J/OL]. Acta Automatica Sinica. (2020-09-21)[2020-09-27]. https://doi.org/10.16383/j.aas.c190866.)

[3] GIBSON J J. The Ecological Approach to Visual Perception: Classic Edition[M]. Hove, East Sussex: Psychology Press, 1979: 89-90.

[4] DODWELL P C. The Lie transformation group model of visual perception[J]. Perception and Psychophysics, 1983, 34(1): 1-16.

[5] LOWE D G. Object recognition from local scale-invariant features [C]// Proceedings of the 7th IEEE International Conference on Computer Vision. Piscataway: IEEE, 1999: 1150-1157.

[6] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]// Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2005: 886-893.

[7] SUNDARAMOORTHI G, PETERSEN P, VARADARAJAN V S, et al. On the set of images modulo viewpoint and contrast changes [C]// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 832-839.

[8] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2012, 25: 1097-1105.

[9] YAO X, NEWSON A, GOUSSEAU Y, et al. A latent transformer for disentangled face editing in images and videos[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 13789-13798.

[10] LECUN Y, JACKEL L D, BOSER B, et al. Handwritten digit recognition: applications of neural network chips and automatic learning[J]. IEEE Communications Magazine, 1989, 27(11): 41-46.

[11] WANG H, ULLAH M M, KLÄSER A, et al. Evaluation of local spatio-temporal features for action recognition[C]// Proceedings of the 2009 British Machine Vision Conference. Durham: BMVA Press, 2009: No. 143.

[12] COURVILLE A, BERGSTRA J, BENGIO Y. A spike and slab restricted Boltzmann machine [C]// Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. New York: JMLR.org, 2011: 233-241.

[13] KOHONEN T. Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map [J]. Biological Cybernetics, 1996, 75(4): 281-291.

[14] HYVÄRINEN A, HOYER P. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces[J]. Neural Computation, 2000, 12(7): 1705-1720.

[15] KAVUKCUOGLU K, RANZATO M, FERGUS R, et al. Learning invariant features through topographic filter maps [C]// Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 1605-1612.

[16] RANZATO M, HINTON G E. Modeling pixel means and covariances using factorized third-order Boltzmann machines[C]// Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2010: 2551-2558.

[17] HIGGINS I, AMOS D, PFAU D, et al. Towards a definition of disentangled representations [EB/OL]. (2018-12-05)[2020-09-27]. https://arxiv.org/pdf/1812.02230.pdf.

[18] DAHL G E, RANZATO M, MOHAMED A R, et al. Phone recognition with the mean-covariance restricted Boltzmann machine [C]// Proceedings of the 23rd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2010: 469-477.

[19] SEIDE F, LI G, YU D. Conversational speech transcription using context-dependent deep neural networks[C]// Proceedings of the 12th Annual Conference of the International Speech Communication Association. Grenoble: ISCA, 2011: 437-440.

[20] CHARTSIAS A, JOYCE T, PAPANASTASIOU G, et al. Disentangled representation learning in cardiac image analysis[J]. Medical Image Analysis, 2019, 58: No. 101535.

[21] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30-42.

[22] 许世斌,高子淑. 基于共性假设的零样本生成模型[J]. 计算机应用与软件, 2020, 37(8):177-181. (XU S B, GAO Z S. A generative model for zero shot learning based on common hypothesis[J]. Computer Applications and Software, 2020, 37(8):177-181.)

[23] 王德文,魏波涛. 基于孪生变分自编码器的小样本图像分类方法[J]. 智能系统学报, 2021, 16(2): 254-262. (WANG D W, WEI B T. A small-sample image classification method based on a Siamese variational auto-encoder [J]. CAAI Transactions on Intelligent Systems, 2021, 16(2): 254-262.)

[24] BOULANGER-LEWANDOWSKI N, BENGIO Y, VINCENT P. Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription[C]// Proceedings of the 29th International Conference on International Conference on Machine Learning. Madison, WI: Omnipress, 2012: 1881-1888.

[25] HAMEL P, LEMIEUX S, BENGIO Y, et al. Temporal pooling and multiscale learning for automatic annotation and ranking of music audio[C]// Proceedings of the 12th International Society for Music Information Retrieval Conference. [S.l.]: ISMIR, 2011: 729-734.

[26] HINTON G E. Learning distributed representations of concepts [M]// MORRIS R G M, Parallel Distributed Processing: Implications for Psychology and Neurobiology. Oxford: Oxford University Press, 1989: 46-61.

[27] BENGIO Y. Neural net language models [J]. Scholarpedia, 2008, 3(1): No. 3881.

[28] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12: 2493-2537.

[29] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions

on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[30] TANG Y B, TANG Y X, ZHU Y Y, et al. A disentangled generative model for disease decomposition in chest X-rays via normal image synthesis[J]. Medical Image Analysis, 2021, 67: No. 101839.

[31] 屠恩美,杨杰. 半监督学习理论及其研究进展概述[J]. 上海交通大学学报, 2018, 52(10):1280-1291. (TU E M, YANG J. A review of semi-supervised learning theories and recent advances [J]. Journal of Shanghai Jiao Tong University, 2018, 52(10): 1280-1291.)

[32] HAO Z F, LV D, LI Z J, et al. Semi-supervised disentangled framework for transferable named entity recognition [J]. Neural Networks, 2021, 135: 127-138.

[33] LI H L, WAN R J, WANG S Q, et al. Unsupervised domain adaptation in the wild via disentangling representation learning [J]. International Journal of Computer Vision, 2021, 129(2): 267-283.

[34] COURVILLE A, BERGSTRA J, BENGIO Y. Unsupervised models of images by spike-and-slab RBMs[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. Madison, WI: Omnipress, 2011: 1145-1152.

[35] DESJARDINS G, COURVILLE A, BENGIO Y. Disentangling factors of variation via generative entangling[EB/OL]. (2012-10-19)[2021-04-22]. https://arxiv.org/pdf/1210.5474.pdf.

[36] HINTON G E, KRIZHEVSKY A, WANG S D. Transforming auto-encoders[C]// Proceedings of the 2011 International Conference on Artificial Neural Networks, LNCS 6791. Berlin: Springer, 2011: 44-51.

[37] MEMISEVIC R, HINTON G E. Learning to represent spatial transformations with factored higher-order Boltzmann machines [J]. Neural Computation, 2010, 22(6): 1473-1492.

[38] TENENBAUM J B, DE SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(5500): 2319-2323.

[39] RIFAI S, VINCENT P, MULLER X, et al. Contractive auto-encoders: explicit invariance during feature extraction [C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. Madison, WI: Omnipress, 2011: 833-840.

[40] TANG Y C, SALAKHUTDINOV R, HINTON G. Tensor analyzers [C]// Proceedings of the 30th International Conference on Machine Learning. New York: JMLR.org, 2013: 163-171.

[41] 卫亮亮. 基于流结构变分推断的深度生成模型研究[D]. 保定:河北大学, 2019. (WEI L L. Research on deep generative models based on variational inference of flow structure [D]. Baoding: Hebei University, 2019.)

[42] BENGIO Y, MESNIL G, DAUPHIN Y, et al. Better mixing via deep representations [C]// Proceedings of the 30th International Conference on Machine Learning. New York: JMLR.org, 2013: 552-560.

[43] REED S, SOHN K, ZHANG Y T, et al. Learning to disentangle factors of variation with manifold interaction [C]// Proceedings of the 31st International Conference on Machine Learning. New York: JMLR.org, 2014: 1431-1439.

[44] BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-127.

[45] LI X Q, CHEN L B, WANG L, et al. SCGAN: disentangled representation learning by adding similarity constraint on generative adversarial nets[J]. IEEE Access, 2019, 7: 147928-147938.

[46] MATHIEU M, ZHAO J B, SPRECHMANN P, et al. Disentangling factors of variation in deep representation using adversarial training [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2016: 5047-5055.

[47] 杨晨曦,左劼,孙频捷. 基于自编码器的零样本学习方法研究进展[J]. 现代计算机, 2020(1):48-52. (YANG C X, ZUO J, SUN P J. Research progress of zero-shot learning method based on autoencoder[J]. Modern Computer, 2020(1):48-52.)

[48] 王路,李寿山. 基于变分自编码器的问题识别方法[J]. 郑州大学学报(理学版), 2019, 51(3):79-84. (WANG L, LI S S. Question detection method based on variational auto-encoder[J]. Journal of Zhengzhou University (Natural Science Edition), 2019, 51(3):79-84.)

[49] 翟正利,梁振明,周炜,等. 变分自编码器模型综述[J]. 计算机工程与应用, 2019, 55(3):1-9. (ZHAI Z L, LIANG Z M, ZHOU W, et al. Research overview of variational auto-encoders models [J]. Computer Engineering and Applications, 2019, 55(3):1-9.)

[50] KINGMA D P, WELLING M. Auto-encoding variational Bayes [EB/OL]. (2014-05-01)[2020-09-27]. https://arxiv.org/pdf/1312.6114.pdf.

[51] HIGGINS I, MATTHEY L, PAL A, et al. β-VAE: learning basic visual concepts with a constrained variational framework[EB/OL]. [2020-09-27]. https://openreview.net/pdf?id=Sy2fzU9gl.

[52] HIGGINS I, SONNERAT N, MATTHEY L, et al. SCAN: learning abstract hierarchical compositional visual concepts [EB/OL]. (2018-06-06) [2020-09-27]. https://arxiv.org/pdf/1707.03389.pdf.

[53] HIGGINS I, CHANG L, LANGSTON V, et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons [J]. Nature Communications, 2021, 12(1): 1-14.

[54] BURGESS C P, HIGGINS I, PAL A, et al. Understanding disentangling in β-VAE[EB/OL]. (2018-04-10)[2020-09-27]. https://arxiv.org/pdf/1804.03599.pdf.

[55] KIM H, MNIH A. Disentangling by factorising[C]// Proceedings of the 35th International Conference on Machine Learning. New York: JMLR.org, 2018: 2649-2658.

[56] PATACCHIOLA M, FOX-ROBERTS P, ROSTEN E. Y-Autoencoders: disentangling latent representations via sequential encoding[J]. Pattern Recognition Letters, 2020, 140: 59-65.

[57] HIGGINS I, PAL A, RUSU A, et al. DARLA: improving zero-shot transfer in reinforcement learning [C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR.org, 2017: 1480-1490.

[58] CHEN R T Q, LI X C, GROSSE R, et al. Isolating sources of disentanglement in VAEs [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018: 2615-2625.

[59] RIDGEWAY K, MOZER M C. Learning deep disentangled embeddings with the F-statistic loss[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018: 185-194.

[60] KUMAR A, SATTIGERI P, BALAKRISHNAN A. Variational inference of disentangled latent concepts from unlabeled observations [EB/OL]. (2018-12-27) [2020-09-27]. https://arxiv.org/pdf/1711.00848.pdf.

[61] IIZUKA S, SIMO-SERRA E, ISHIKAWA H. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification [J]. ACM Transactions on Graphics, 2016, 35(4): No. 110.

[62] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 2650-2658.

[63] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5967-5976.

[64] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2223-2232.

[65] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation [C]// Proceedings of the 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS 9351. Cham: Springer, 2015: 234-241.

[66] MA L, SUN Q, GEORGOULIS S, et al. Disentangled person image generation [C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 99-108.

[67] BARROW H G, TENENBAUM J M. Recovering intrinsic scene characteristics from images [M]// HANSON A, RISEMAN E, Computer Vision Systems. New York: Academic Press, 1978: 3-26.

[68] Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[69] GONZALEZ-GARCIA A, WEIJER J V D, BENGIO Y. Image-to-image translation for cross-domain disentanglement [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018: 1294-1305.

[70] ZHU J Y, ZHANG R, PATHAK D, et al. Toward multimodal image-to-image translation [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 465-476.

[71] LEE H Y, TSENG H Y, MAO Q, et al. DRIT++: diverse image-to-image translation via disentangled representations [J]. International Journal of Computer Vision, 2020, 128(10/11): 2402-2417.

[72] 白静,田栋文,张霖,等. 跨域变分对抗自编码器[J]. 计算机辅助设计与图形学学报, 2020, 32(9):1402-1410. (BAI J, TIAN D W, ZHANG L, et al. Cross-domain variational adversarial autoencoder[J]. Journal of Computer-Aided Design and Graphics, 2020, 32(9):1402-1410.)

[73] CHIAPPA S, RACANIERE S, WIERSTRA D, et al. Recurrent environment simulators [EB/OL]. (2017-04-19) [2020-09-27]. https://arxiv.org/pdf/1704.02254.pdf.

[74] AGRAWAL P, NAIR A, ABBEEL P, et al. Learning to poke by poking: experiential learning of intuitive physics [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2016: 5092-5100.

[75] DENTON E, BIRODKAR V. Unsupervised learning of disentangled representations from video [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 4417-4426.

[76] 徐思. 基于对抗自编码器的半监督分类模型研究[D]. 青岛: 青岛大学, 2019. (XU S. Research on semi-supervised classification model based on adversarial auto-encoder [D]. Qingdao: Qingdao University, 2019.)

[77] SUN P F, SU X, GUO S Q, et al. Cycle representation-disentangling network: learning to completely disentangle spatial-temporal features in video [J]. Applied Intelligence, 2020, 50(12): 4261-4280.

[78] ZHU J Y, ZHANG Z T, ZHANG C K, et al. Visual object networks: image generation with disentangled 3D representations [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018: 118-129.

[79] LOCATELLO F, BAUER S, LUCIC M, et al. Challenging common assumptions in the unsupervised learning of disentangled representations [C]// Proceedings of the 36th International Conference on Machine Learning. New York: JMLR.org, 2019: 4114-4124.

[80] LOCATELLO F, BAUER S, LUCIC M, et al. A sober look at the unsupervised learning of disentangled representations and their evaluation [J]. Journal of Machine Learning, 2020, 21: 1-62.

[81] VAN STEENKISTE S, LOCATELLO F, SCHMIDHUBER J, et al. Are disentangled representations helpful for abstract visual reasoning? [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2019: 14245-14258.

[82] DO K, TRAN T. Theory and evaluation metrics for learning disentangled representations [EB/OL]. (2020-02-04) [2020-09-27]. https://arxiv.org/pdf/1908.09961.pdf.

This work is partially supported by the National Natural Science Foundation of China (61972183, 61602215), the Director Foundation of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data.

CHENG Keyang, born in 1982, Ph. D., professor. His research interests include computer vision, pattern recognition.

MENG Chunyun, born in 1994, M. S. candidate. His research interests include computer vision, pattern recognition.

WANG Wenshan, born in 1994, M. S. Her research interests include statistical analysis, machine learning.

SHI Wenxi, born in 1988, Ph. D. His research interests include big data analysis, smart security.

ZHAN Yongzhao, born in 1962, Ph. D., professor. His research interest include multimedia, artificial intelligence.