第1章

表征学习

Liang Zhao、Lingfei Wu、Peng Cui 和 Jian Pei[①]

摘要

在本章中，我们将首先介绍什么是表征学习以及为什么需要表征学习。在表征学习的各种方式中，本章重点讨论的是深度学习方法：那些由多个非线性变换组成的方法，目的是产生更抽象且最终更有用的表征。接下来，我们将总结不同领域的表征学习技术，重点是不同数据类型的独特挑战和模型，包括图像、自然语言、语音信号和网络等。最后，我们将总结本章的内容，并提供基于互信息的表征学习的延伸阅读材料——一种最近出现的通过无监督学习的表征技术。

1.1 导读

机器学习技术的有效性在很大程度上不仅依赖于算法本身的设计，而且依赖于良好的数据表征（特征集）。由于缺少一些重要信息、包含不正确信息或存在大量冗余信息，无效数据表征会导致算法在处理不同任务时表现不佳。表征学习的目标是从数据中提取足够但最少的信息。传统上，该目标可以通过先验知识以及基于数据和任务的领域专业知识来实现，这也被称为特征工程。历史上，在部署机器学习和许多其他人工智能算法时，很大一部分人力需要投到预处理过程和数据转换中。更具体地说，特征工程是利用人类的聪明才智和现有知识的一种方式，旨在从数据中提取并获得用于机器学习任务的判别信息。例如，政治学家可能定义一个关键词列表用作社交媒体文本分类器的特征，以检测那些关于社会事件的文本。对于语音转录识别，人们可以通过相关操作（如傅里叶变换等）从原始声波中提取特征。尽管多年来特征工程得到了广泛应用，但其缺点也很突出，包括：（1）通常需要领域专家的密集劳动，这是因为特征工程可能需要模型开发者和领域专家之间紧密而广泛的合作；（2）不完整的和带有偏见的特征提取。具体来说，不同领域专家的知识限制了所提取特征的容量和判别能力。此外，在许多人类知识有限的领域，提取什么特征本身就是领域专家的一个开放性问题，如癌症早期预测。为了避免这些缺点，使得学习算法不

① Liang Zhao
Department of Computer Science，Emory University，E-mail：liang.zhao@emory.edu
Lingfei Wu
Pinterest，E-mail：lwu@email.wm.edu
Peng Cui
Department of Computer Science，Tsinghua University，E-mail：cuip@tsinghua.edu.cn
Jian Pei
Department of Computer Science，Simon Fraser University，E-mail：jpei@cs.sfu.ca

那么依赖特征工程，一直是机器学习和人工智能领域的一个非常理想的目标，由此可以快速构建新的应用，并有望更有效地解决问题。

表征学习的技术见证了从传统表征学习到更先进表征学习的发展与演变。传统的表征学习方法属于“浅层”模型，旨在学习数据转换，使其在建立分类器或其他预测器时更容易提取有用的信息，如主成分分析（Principal Component Analysis，PCA）（Wold et al，1987）、高斯马尔可夫随机场（Gaussian Markov Random Field，GMRF）（Rue and Held，2005）以及局部保持投影（Locality Preserving Projections，LPP）（He and Niyogi，2004）。基于深度学习的表征学习则由多个非线性变换组成，目的是产生更抽象且更有用的表征。为了介绍更多的最新进展并聚焦本书的主题，本节主要关注基于深度学习的表征学习，具体可以分为以下三种类型：（1）监督学习，需要通过大量的标记数据训练深度学习模型。给定训练良好的网络，最后一个全连接层之前的输出总是被用作输入数据的最终表征。（2）无监督学习（包括自监督学习），有利于分析没有相应标签的输入数据，旨在学习数据的潜在固有结构或分布，通过代理任务可以从大量无标签数据中探索监督信息。基于这种方式构建的监督信息可以训练深度神经网络，从而为未来下游任务提取有意义的表征。（3）迁移学习（Transfer Learning，TL），涉及利用任何知识资源（如数据、模型、标签等）增加模型对目标任务的学习和泛化能力。迁移学习囊括不同的场景，如多任务学习（Multi-Task Learning，MTL）、模型适应、知识迁移、协变量偏移等。其他重要的表征学习方法还有强化学习、小样本学习和解耦表征学习等。

定义什么是好的表征很重要。正如 Bengio（2008）所定义的那样，表征学习是关于学习数据的（底层）特征。在建立分类器或其他预测器时，基于表征更容易提取有用的信息。因此，对所学表征的评价与其在下游任务中的表现密切相关。例如，在基于生成模型的数据生成任务中，对于观察到的输入，好的表征往往能够捕捉到潜在解释因素的后验分布；而对预测任务来说，好的表征能够捕捉到输入数据的最少但足够的信息来正确预测目标标签。除从下游任务的角度进行评价以外，还可以基于好的表征可能具有的一般属性进行评价，如平滑性、线性、捕捉多个解释性的或因果性的因素、在不同任务之间保持共同因素以及简单的因素依赖性等。

1.2 不同领域的表征学习

在本节中，我们将总结表征学习在 4 个不同的代表性领域的发展状况：（1）图像处理；（2）语音识别；（3）自然语言处理；（4）网络分析。对于每个研究领域的表征学习，我们将考虑一些推动该领域研究的基本问题。具体来说，是什么让一个表征比另一个表征更好，以及应该如何计算表征？为什么表征学习在该领域很重要？另外，学习好的表征的适当目标是什么？我们还将分别从监督表征学习、无监督表征学习和迁移学习三方面介绍相关的典型方法及其发展状况。

1.2.1 用于图像处理的表征学习

图像表征学习是理解各种视觉数据（如照片、医学图像、文件扫描和视频流等）的语义的一个基本问题。通常情况下，图像处理中的图像表征学习的目标是弥合像素数据和图像语义之间的语义差距。图像表征学习已经成功解决了现实世界里的许多问题，包括但不

限于图像搜索、面部识别、医学图像分析、照片处理和目标检测等。

近年来，我们见证了图像表征学习从手工特征工程到通过深度神经网络模型自动处理的快速发展过程。传统上，图像的模式是由人们基于先验知识借助手工特征提取的。例如，Huang et al（2000）从笔画中提取了字符的结构特征，然后用它们识别手写字符。Rui（2005）采用形态学方法改善了字符的局部特征，然后使用 PCA 提取字符的特征。然而，所有这些方法都需要手动从图像中提取特征，因此相关的预测表现强烈依赖于先验知识。在计算机视觉领域，由于特征向量具有高维度，手动提取特征是非常烦琐和不切实际的。因此，能够从高维视觉数据中自动提取有意义的、隐藏的、复杂的模式，这样的图像表征学习是必要的。基于深度学习的图像表征学习是以端到端的方式学习的，只要训练数据的质量足够高、数量足够多，其在目标应用中的表现就比手动制作的特征要好得多。

用于图像处理的监督表征学习。在图像处理领域，监督学习算法，如卷积神经网络（Convolution Neural Network，CNN）和深度信念网络（Deep Belief Network，DBN），被普遍应用于解决各种任务。最早的基于深度监督学习的成果之一是在 2006 年提出的（Hinton et al，2006），它专注于处理 MNIST 数字图像分类问题，其表现优于最先进的支持向量机（Support Vector Machine，SVM）。自此，深度卷积神经网络（ConvNets）表现出惊人的性能，这在很大程度上取决于它们的平移不变性、权重共享和局部模式捕获等特性。为了提高网络模型的容量，人们开发了不同类型的网络架构，而且收集的数据集越来越大。包括 AlexNet（Krizhevsky et al，2012）、VGG（Simonyan and Zisserman，2014b）、GoogLeNet（Szegedy et al，2015）、ResNet（He et al，2016a）和 DenseNet（Huang et al，2017a）等在内的各种网络以及 ImageNet、OpenImage 等大规模数据集都可以用于训练深层的卷积神经网络。凭借复杂的架构和大规模数据集，卷积神经网络在各种计算机视觉任务中不断超越之前最先进的技术。

用于图像处理的无监督表征学习。在图像数据集和视频数据集中，大规模数据集的收集和标注都很耗时且昂贵。例如，ImageNet 包含大约 130 万张有标签的图像，涵盖 1 000 个类别，每张图像都由人工标注了一个类别标签。为了减少大量的人工标注工作，人们提出了许多用于从大规模未标注的图像或视频中学习视觉特征的无监督方法，而无须任何人工标注。一种流行的解决方案是提出各种代理任务供模型解决，模型则通过学习代理任务的目标函数进行训练，并通过这个过程学习特征。针对无监督学习，人们提出了各种代理任务，包括灰度图像着色（Zhang et al，2016d）和图像修复（Pathak et al，2016）。在无监督训练阶段，需要设计供模型解决的预定义的代理任务，代理任务的伪标签是根据数据的一些属性自动生成的，然后根据代理任务的目标函数训练模型。当使用代理任务进行训练时，深度神经网络模型的浅层部分侧重于低层次的一般特征，如角落、边缘和纹理等，而深层部分则侧重于高层次的特定任务特征，如物体、场景等。因此，用预先定义的代理任务训练的模型可以通过学习内核来捕捉低层次和高层次的特征，这些特征对其他下游任务是有帮助的。在无监督训练结束后，这种在预训练模型中学习到的视觉特征便可以进一步迁移到下游任务中（特别是在只有相对较少的数据时），以提高表现并克服过拟合。

用于图像处理的迁移学习。在现实世界的应用中，由于人工标注的成本很高，可能并非总是可以获得足够的属于相同特征空间或测试数据分布的训练数据。迁移学习通过模仿人类视觉系统，在给定领域（即目标领域）执行新任务时，利用了其他相关领域（即源领域）的足够数量的先验知识。在迁移学习中，针对目标领域和源领域，训练集和测试集都

可以起作用。大多数情况下，一个迁移学习任务只有一个目标领域，但可以存在一个或多个源领域。用于图像处理的迁移学习技术分为特征表征知识迁移和基于分类器的知识迁移两种。具体来说，特征表征知识迁移利用一组提取的特征将目标领域映射到源领域，这样可以显著减少目标领域和源领域之间的数据差异，从而提高目标领域的任务性能。基于分类器的知识迁移则通常有一个共同的特点，也就是将学到的源领域模型作为先验知识，用于与训练样本一起学习目标模型。基于分类器的知识迁移不是通过提高实例的表征来最小化跨领域的不相似性，而是通过提供的两个领域的训练集和学习的模型来学习另一个新的模型，进而使目标领域的泛化误差最小。

用于图像处理的其他表征学习技术。其他类型的表征学习技术也被经常用于图像处理，如强化学习和半监督学习。例如，可以尝试在一些任务中使用强化学习，如图像描述（Liu et al，2018a；Ren et al，2017）以及图像编辑（Kosugi and Yamasaki，2020），其中的学习过程可被形式化为基于策略网络的一系列行动。

1.2.2 用于语音识别的表征学习

如今，现实生活里的各种应用中和设备上已经广泛集成或开发了语音接口或系统。像Siri[①]、Cortana[②]和谷歌语音搜索[③]这样的服务已经成为人们生活的一部分，被数百万用户使用。对语音识别和分析进行探索的初衷是希望机器能够提供人机交互服务。60 多年来，使机器能够理解人类语音、识别说话者和检测人类情感的研究目标吸引了越来越多研究人员的注意力，涉及的研究领域包括自动语音识别（Automatic Speech Recognition，ASR）、说话者识别（Speaker Recognition，SR）和说话者情感识别（Speaker Emotion Recognition，SER）等。

分析和处理语音一直是机器学习算法的一个关键应用。传统上，关于语音识别的研究认为，设计手工声学特征的任务与设计有效模型以完成预测和分类决策的任务是彼此独立的两个不同问题。这种方法有两个主要缺点。首先，如前所述，特征工程比较麻烦，涉及人类的先验知识；其次，设计的特征可能不是针对特定语音识别任务的最佳选择。这促使语音社群尝试使用表征学习技术的最新成果，以自动学习输入信号的中间表征，更好地适应将要面临的任务，进而提高性能。在所有这些成功的尝试中，基于深度学习的语音表征发挥了重要作用。我们在语音技术中利用表征学习技术的原因之一在于语音数据与二维图像数据有如下根本区别：图像可以作为一个整体或块进行分析，但语音必须按顺序格式，以捕捉时间依赖性和模式。

用于语音识别的监督表征学习。在语音识别和分析领域，监督表征学习得到了广泛应用，其中的特征表征是通过标签信息在数据集上学习的。例如，受限玻尔兹曼机（Restricted Boltzmann Machine，RBM）（Jaitly and Hinton，2011；Dahl et al，2010）和深度信念网络（Cairong et al，2016；Ali et al，2018）通常用于从语音中学习特征，以处理不同的任务，包括 ASR、SR 和 SER。2012 年，微软发布了 MAVIS（Microsoft Audio Video Indexing Service）语音系统的新版本，该系统基于依赖上下文的深度神经网络（Seide et al，2011）。与基于高斯混合的传统模型相比，开发人员成功地将 4 个主要基准数据集上的单词错误率降低了约 30%（例如，在 RT03S

① Siri 是 iOS 系统内置的一款人工智能助理软件。
② Cortana 是微软开发的智能个人助理，被称为“全球首个跨平台的智能个人助理”。
③ 谷歌语音搜索是谷歌的一款产品，用户可以通过对着手机或计算机说话来使用谷歌语音搜索。工作过程是首先利用服务器识别设备上的内容，然后根据识别结果搜索信息。

上从 27.4%降至 18.5%）。卷积神经网络是另一种流行的监督模型，被广泛用于诸如语音和说话人识别等任务中的语音信号特征学习（Palaz et al，2015a，b）和 SER（Latif et al，2019；Tzirakis et al，2018）。此外，人们发现 LSTM（或 GRU）可以学习局部和长期依赖，从而帮助 CNN 从语音中学习更多有用的特征（Dahl et al，2010）。

用于语音识别的无监督表征学习。利用大型无标签数据集进行无监督表征学习是语音识别的一个活跃领域。在语音分析中，这种技术支持利用实际可用的无限量的无标签语料来学习良好的中间特征表征，这些中间特征表征可用于提高各种下游监督学习语音识别任务或语音信号合成任务的表现。在 ASR 和 SR 任务中，大多数工作是基于变分自编码器（Variational AutoEncoder，VAE）的，其中的生成模型和推理模型是联合学习的，这使得它们能够从观察到的语音数据中捕获潜在的表征（Chorowski et al，2019；Hsu et al，2019，2017）。例如，Hsu et al（2017）提出了分层 VAE，旨在没有任何监督的情况下从语音中捕捉可以解释和解耦的表征。其他自编码架构，如降噪自编码器（Denoised AutoEncoder，DAE），在以无监督方式寻找语音表征方面非常有前途，尤其是针对嘈杂语音的识别（Feng et al，2014；Zhao et al，2015）。除上述成果以外，最近，对抗性学习（Adversarial Learning，AL）正在成为学习无监督语音表征的有力工具，如生成对抗网络（Generative Adversarial Net，GAN）。GAN 至少涉及一个生成器和一个判别器，前者试图生成尽可能真实的数据来混淆后者，后者则尽力试图去除混淆。因此，生成器和判别器都能够以对抗方式进行训练和反复改进，从而产生更多具有判别性和鲁棒性的特征。其中，GAN（Chang and Scherer，2017；Donahue et al，2018）、对抗性自编码器（AAE）（Sahu et al，2017）不仅在 ASR 的语音建模中，而且在 SR 和 SER 的语音建模中正变得越来越流行。

用于语音识别的迁移学习。迁移学习（Transfer Learning，TL）囊括不同的场景，如 MTL、模型自适应、知识迁移、协变量偏移等。在语音识别领域，表征学习在 TL 的这些场景中得到了极大发展，包括领域自适应、多任务学习和自主学习等。就域适应而言，语音数据是典型的异质数据。因此，源域数据和目标域数据的概率分布之间总是存在不匹配的情况。为了在现实生活中构建更强大的语音相关应用系统，我们通常在深度神经网络的训练解决方案中应用域适应技术，以学习能够显式最小化源域数据和目标域数据分布之间差异的表征（Sun et al，2017；Swietojanski et al，2016）。就 MTL 而言，表征学习可以成功地提高语音识别的性能，而不需要上下文语音数据，这是因为语音包含用作辅助任务的多维信息（如消息、说话者、性别或情感等）。例如，在 ASR 任务中，通过将 MTL 与不同的辅助任务（包括性别、说话者适应、语音增强等）结合使用，研究表明，为不同任务学习的共享表征可以作为声学环境的补充信息，并表现出较低的单词错误率（Word Error Rate，WER）（Parthasarathy and Busso，2017；Xia and Liu，2015）。

用于语音识别的其他表征学习技术。除上述三类用于语音识别的表征学习技术以外，还有一些其他的表征学习技术受到广泛关注，如半监督学习和强化学习（Reinforcement Learning，RL）。例如，在 ASR 任务中，半监督学习主要用于解决缺乏足够训练数据的问题，这可以通过创建特征前端（Thomas et al，2013）、使用多语言声学表征（Cui et al，2015）或从大型未配对数据集中提取中间表征（Karita et al，2018）来实现。RL 在语音识别领域也受到广泛关注，并且已经有多种方法可以对不同的语音问题进行建模，包括对话建模和优化（Levin et al，2000）、语音识别（Shen et al，2019）和情感识别（Sangeetha and Jayasankar，2019）。

1.2.3 用于自然语言处理的表征学习

除语音识别以外，表征学习还有许多其他自然语言处理（Natural Language Processing，NLP）方面的应用，如文本表征学习。谷歌图像搜索基于 NLP 技术利用大量数据把图像和查询映射到了同一空间（Weston et al，2010）。一般来说，表征学习在 NLP 中的应用有两种类型。在其中一种类型中，语义表征（如词嵌入）是在预训练任务中训练的（或直接由专家设计），然后被迁移到目标任务的模型中。语义表征通过语言建模目标进行训练，并作为其他下游 NLP 模型的输入。在另一种类型中，语义表征暗含在深度学习模型中，并直接以端到端的方式更好地实现目标任务。例如，许多 NLP 任务希望在语义上合成句子表征或文档表征，如情感分类、自然语言推理和关系提取等需要句子表征的任务。

传统的 NLP 任务严重依赖特征工程，这需要精心的设计和大量的专业知识。表征学习（特别是基于深度学习的表征学习）正成为近年来 NLP 最重要的技术。首先，NLP 通常关注多层次的语言条目，包括字符、单词、短语、句子、段落和文档等。表征学习能够在统一的语义空间中表征这些多层次语言条目的语义，并在这些语言条目之间建立复杂的语义依赖模型。其次，可以在同一输入上执行各种 NLP 任务。给定一个句子，我们可以执行多个任务，如单词分割、命名实体识别、关系提取、共指链接和机器翻译等。在这种情况下，为多个任务建立一个统一的输入表征空间将更加有效和稳健。最后，可以从多个领域收集自然语言文本，包括新闻文章、科学文章、文学作品、广告以及在线用户生成的内容，如产品评价和社交媒体等。此外，也可以从不同的语言中收集这些文本，如英语、汉语、西班牙语、日语等。与传统的 NLP 系统必须根据每个领域的特点设计特定的特征提取算法相比，表征学习能够使我们从大规模领域数据中自动构建表征，甚至在来自不同领域的这些语言之间建立桥梁。鉴于 NLP 表征学习在减少特征工程和性能改进方面的这些优势，许多研究人员致力于开发高效的表征学习算法，尤其是用于深度学习的 NLP 方法。

用于 NLP 的监督表征学习。近年来，用于 NLP 的监督学习设定下的深度神经网络中首先出现的是分布式表征学习，然后是 CNN 模型，最后是 RNN 模型。早期，Bengio 等人首先在统计语言建模的背景下开发了分布式表征学习，Bengio et al（2008）将其称为神经网络语言模型，该模型用于为每个词学习一个分布式表征（即词嵌入）。之后，我们需要一个从构成词或 n 元文法中提取更高层次特征的有效特征函数。鉴于 CNN 在计算机视觉和语音处理任务中出色的表现，CNN 顺理成章地被选中。CNN 有能力从输入的句子中提取突出的 n 元文法特征，从而为下游任务创建句子的信息潜在语义表征。这一领域由 Collobert et al（2011）和 Kalchbrenner et al（2014）开创，它使得基于 CNN 的网络在随后的文献中被广泛引用。通过在隐藏层中加入循环（Mikolov et al，2011a）（如 RNN），神经网络语言模型得到改进，其不仅在复杂度（预测正确下一个单词的平均负对数似然的指数）方面，而且在语音识别的误码率方面，能够击败最先进的模型（平滑的 n 元文法模型）。RNN 则采用了处理顺序信息的思路。之所以采用术语“循环”，是因为神经网络语言模型对序列中的每个词条都会进行相同的计算，并且每一步都依赖于先前的计算和结果。一般来说，可通过将词条逐个送入循环单元来生成一个固定大小的向量以表征一个序列。在某种程度上，RNN 对以前的计算具有“记忆”，支持在当前处理的任务中使用这些信息。这种模型自然适用于许多 NLP 任务，如语言建模（Mikolov et al，2010，2011b）、机器翻译（Liu et al，2014；Sutskever et al，2014）以及图像描述（Karpathy and Fei-Fei，2015）。

用于 NLP 的无监督表征学习。无监督学习（包括自监督学习）在 NLP 领域取得了巨大成功，这是因为纯文本本身含有丰富的语言知识和模式。例如，在大多数基于深度学习的 NLP 模型中，句子中的单词首先通过 word2vec（Mikolov et al，2013b）、GloVe（Pennington et al，2014）和 BERT（Devlin et al，2019）等技术被映射到相关的嵌入，然后被送入网络。不过，我们没有用于学习这些词嵌入的人工标注的“标签”。为了获得神经网络所需的训练目标，有必要从现有数据中产生内在的“标签”。语言建模是典型的无监督学习任务，可以构建单词序列的概率分布，而无须人工标注。基于分布假设，使用语言建模的目标可以获得编码单词语义的隐藏表征。在 NLP 中，另一个典型的无监督学习模型是自编码器，由降维（编码）阶段和重建（解码）阶段组成。例如，循环自编码器（其囊括具有 VAE 的循环网络）已经在全句转述检测中超越了最先进的技术，Socher et al（2011）将用于评估副词检测效果的 F1 分数几乎翻了一番。

用于 NLP 的迁移学习。近年来，在 NLP 领域，顺序迁移学习模型和架构的应用印证了迁移学习方法的快速发展，这些方法在广泛的 NLP 任务中极大改善了相关技术水平。在领域适应方面，顺序迁移学习包括两个阶段：首先是预训练阶段，主要包括在源任务或领域中学习一般的表征；其次是适应阶段，主要包括将学到的知识应用于目标任务或领域。NLP 中的领域适应可以分为以模型为中心、以数据为中心和混合方法三种。以模型为中心的方法旨在增强特征空间以及改变损失函数、结构或模型参数（Blitzer et al，2006）。以数据为中心的方法专注于数据方面，涉及伪标签（或自举），其中只有少量的类别在源数据集和目标数据集之间共享（Abney，2007）。混合方法是由以数据和模型为中心的模型建立的。同样，NLP 在多任务学习方面也取得了很大的进展，不同的 NLP 任务可以具有更好的文本表达。例如，基于卷积架构（Collobert et al，2011）开发的 SENNA 系统在语言建模、词性标签、分块、命名实体识别、语义角色标记和句法解析等任务中共享表征。在这些任务上，SENNA 接近甚至有时超过最先进的水平，同时相比传统的预测器在结构上更简单，处理速度更快。此外，学习词嵌入可以与学习图像表征相结合，从而将文本和图像关联起来。

用于 NLP 的其他表征学习技术。在 NLP 任务中，当一个问题变得比较复杂时，就需要领域专家提供更多的知识来标注细粒度任务的训练实例，这将增加标注数据的成本。因此，有时需要通过（非常）少的标注数据来有效地开发模型或系统。当每个类别只有一个或几个标注的实例时，问题就变成单样本/少样本学习问题。少样本学习问题源于计算机视觉，最近才开始应用于 NLP。例如，研究人员已经探索了少样本关系提取（Han et al，2018），其中每个关系都有几个标注实例以及并行语料库规模有限的低资源机器翻译（Zoph et al，2016）。

1.2.4 用于网络分析的表征学习

除文本、图像和声音等常见数据类型以外，网络数据是另一种重要的数据类型。在现实世界的大规模应用中，网络数据无处不在，从虚拟网络（如社交网络、引用网络、电信网络等）到现实网络（如交通网络、生物网络等）。网络数据在数学上可以表述为图，其中的顶点（节点）及其之间的关系共同表征了网络信息。网络和图是非常强大和灵活的数据表述方式，有时我们甚至可以把其他数据类型（如文本和图像）看作它们的特例。例如，图像可以认为是具有 RGB 属性的节点网格，它们是特殊类型的图；而文本也可以组织成顺序的、树状的或图结构的信息。因此，总的来说，网络的表征学习已被广泛认为是一项有

前途但更具挑战性的任务，需要我们推动和促进许多针对图像、文本等开发的技术的发展。除网络数据固有的高复杂性以外，考虑到现实世界中的许多网络规模庞大，拥有从几百到几百万甚至几十亿个顶点，网络的表征学习的效率也是一个重要的问题。分析信息网络在许多学科的各种新兴应用中具有关键作用。例如，在社交网络中，将用户分类为有意义的社会群体对许多重要的任务是有用的，如用户搜索、有针对性的广告和推荐等；在通信网络中，检测群落结构可以帮助机构更好地理解谣言的传播过程；在生物网络中，推断蛋白质之间的相互作用可以促进研究治疗疾病的新方法。然而，对这些网络的高效和有效分析在很大程度上依赖于网络的良好表征。

传统的网络数据特征工程通常侧重于通过图层面（如直径、平均路径长度和聚类系数）、节点层面（如节点度和中心度）或子图层面（如频繁子图和图主题）获得一些预定义的直接特征。虽然这些手动打造的、定义明确的、数量有限的特征描述了图的几个基本方面，但却抛弃了那些不能被它们覆盖的模式。此外，现实世界中的网络现象通常是高度复杂的，需要通过由这些预定义特征组成的、复杂的、未知的组合来描述，也可能无法用任何现有的特征来描述。另外，传统的图特征工程通常涉及昂贵的计算以及具有超线性或指数级的复杂性，这些问题往往使得许多网络分析任务的计算成本高企，难以在大规模网络中使用。例如，在处理群落检测任务时，经典的方法涉及计算矩阵的谱分解，其时间复杂度至少与顶点数量成四次方关系。这种计算成本使得算法难以扩展到具有数百万个顶点的大规模网络。

最近，网络表征学习（Network Representation Learning，NRL）引起了很多人的研究兴趣。NRL 旨在学习潜在的、低维的网络顶点表征，同时保留网络拓扑结构、顶点内容和其他侧面信息。在学习新的顶点表征之后，通过对新的表征空间应用传统的基于向量的机器学习算法，就可以轻松、有效地处理网络分析任务。早期与网络表征学习相关的工作可以追溯到 21 世纪初，当时研究人员提出了将图嵌入算法作为降维技术一部分的观点。给定一组独立且分布相同的数据点作为输入，图嵌入算法首先计算成对数据点之间的相似性，以构建一个亲和图，如 k 近邻图，然后将这个亲和图嵌入一个具有更低维度的新空间。然而，图的嵌入算法主要是为降维设计的，其时间复杂度通常与顶点的数量有关，至少是平方复杂度。

自 2008 年以来，大量的研究工作转向开发直接为复杂信息网络设计的有效且可扩展的表征学习技术。许多网络表征学习算法（Perozzi et al，2014；Yang et al，2015b；Zhang et al，2016b；Manessi et al，2020）已经被提出来并嵌入现有的网络，这些算法在各种应用中表现良好，它们通过将网络嵌入一个潜在的低维空间而保留了结构相似性和属性相似性，由此产生的紧凑、低维的矢量表征可以作为任何基于矢量的机器学习算法的特征，这为我们在新的矢量空间中轻松、有效地处理各种网络分析任务铺平了道路，如节点分类（Zhu et al，2007）、链接预测（Lüand Zhou，2011）、聚类（Malliaros and Vazirgiannis，2013）、网络合成（You et al，2018b）等。本书后续各章将对网络表征学习进行系统而全面的介绍。

1.3 小结

表征学习是目前非常活跃和重要的一个领域，它在很大程度上影响着机器学习技术的有效性。表征学习是指学习数据的表征，使其在建立分类器或其他预测器时更容易提取有

用的、具有鉴别性的信息。当前，在各种学习表征的算法中，深度学习算法已经在诸多领域得到广泛应用。在这些领域，深度学习算法可以基于大量复杂的高维数据，高效且自动地学习好的表征。我们对一个表征做出的评价与其在下游任务中的表现密切相关。一般来说，好的表征除有一些常见属性（如平滑性、线性、离散性）以外，通常还会有一些特殊的属性用于捕捉多个解释性的或因果性的因素。

在本章中，我们总结了不同领域的表征学习技术，重点介绍了不同领域的独特挑战和模型，包括图像、自然语言和语音信号的处理。这些领域都出现了许多基于深度学习的表征技术，可分为监督学习、无监督学习、迁移学习、解耦表征学习、强化学习等不同类别。此外，我们还简要介绍了网络上的表征学习及其与图像、文本和语音的关系，这些内容我们将在后续章节中详细阐述。