计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(5):1−20 [doi: 10.15888/j.cnki.csa.008463]

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

零样本图学习综述①

支瑞聪[1,2], 万　菲[1,2], 张德政[1,2]

[1](北京科技大学 计算机与通信工程学院, 北京 100083)
[2](材料领域知识工程北京市重点实验室, 北京 100083)
通信作者: 支瑞聪, E-mail: zhirc_research@126.com

摘　要: 深度学习方法的提出使得机器学习研究领域得到了巨大突破, 但是却需要大量的人工标注数据来辅助完成. 在实际问题中, 受限于人力成本, 许多应用需要对从未见过的实例类别进行推理判断. 为此, 零样本学习 (zero-shot learning, ZSL) 应运而生. 图作为一种表示事物之间联系的自然数据结构, 目前在零样本学习中受到了越来越多的关注. 本文对零样本图学习方法进行了系统综述. 首先概述了零样本学习和图学习的定义, 并总结了零样本学习现有的解决方案思想. 然后依据图的不同利用方式对目前零样本图学习的方法体系进行了分类. 接下来讨论了零样本图学习所涉及到的评估准则和数据集. 最后指明了零样本图学习进一步研究中需要解决的问题以及未来可能的发展方向.

关键词: 零样本学习; 图学习; 跨模态学习; 属性; 词向量; 流形对齐; 深度学习; 图像识别

引用格式: 支瑞聪,万菲,张德政.零样本图学习综述.计算机系统应用,2022,31(5):1–20. http://www.c-s-a.org.cn/1003-3254/8463.html

Overview on Graph-based Zero-shot Learning

ZHI Rui-Cong[1,2], WAN Fei[1,2], ZHANG De-Zheng[1,2]

[1](School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)
[2](Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China)

Abstract: Although the deep learning method has made a huge breakthrough in machine learning, it requires a large amount of manual work for data annotation. Limited by labor costs, however, many applications are expected to reason and judge the instance labels that have never been encountered before. For this reason, zero-shot learning (ZSL) came into being. As a natural data structure that represents the connection between things, the graph is currently drawing more and more attention in ZSL. Therefore, this study reviews the methods of graph-based ZSL systematically. Firstly, the definitions of ZSL and graph learning are outlined, and the ideas of existing solutions for ZSL are summarized. Secondly, the current ZSL methods are classified according to different utilization ways of graphs. Thirdly, the evaluation criteria and datasets concerning graph-based ZSL are discussed. Finally, this study also specifies the problems to be solved in further research on graph-based ZSL and predicts the possible directions of its future development.

Key words: zero-shot learning (ZSL); graph learning; cross-modal learning; attribute; word vector; manifold alignment; deep learning; image recognition

在机器学习领域中, 有两种典型的学习范式. 一种是有监督学习, 指的是从标签化数据集中推断出对应函数映射的机器学习任务. 它通过对输入数据和输出数据之间的关系进行建模, 生成一个从实例对象特征到实例标签的映射, 并能够将这个映射应用到其他具有相同标签集的数据集上. 另一种是无监督学习, 其目

① 基金项目: 国家自然科学基金面上项目 (61673052); 中央高校基本科研业务费专项资金 (FRF-TP-20-10B, FRF-GF-19-010A, FRF-IDRY-19-011)
收稿时间: 2021-07-14; 修改时间: 2021-08-18; 采用时间: 2021-08-31; csa 在线出版时间: 2022-02-21

的在于找到一个函数映射对数据进行分类, 以了解数据分布. 区别于有监督学习, 无监督学习的特点是数据没有标签信息, 函数映射针对提供的输入范例找出潜在的聚类规则. 在训练结束后, 这个映射也可以用到新的实例上, 得到测试实例所属的聚簇. 但无监督学习并不能给出实例所属的具体类别, 并且由于缺少标签信息的监督, 难以有效的评估其聚类效果.

近年来, 深度学习作为一种特殊的机器学习方式取得了巨大进展, 并在机器学习的各个领域都有了很大的突破. 然而, 深度学习在有监督学习范式上能够产生作用的重要原因之一在于其需要海量的训练标注数据, 这些标注往往需要耗费人工巨大的时间和精力. 同时, 测试数据集标签必须和训练数据集完全一致, 即有监督学习所产生的映射只能处理同类别对象的数据, 而无法迁移到其他类别的判定上. 无监督学习虽然不需要标签监督过程, 避免了标注的复杂性和专业性的限制, 但却并不能够提供实例的类别, 这和实践中的期望是相违背的. 更重要的是, 由于自然界中的数据往往是长尾分布的, 即大多数类别都不具备足够且合适的训练实例, 因此常常会出现训练实例的类别未能覆盖测试类别的情况.

为了解决监督学习与非监督学习的限制, 受到人类学习行为的启发, 研究人员提出了零样本学习 (zero-shot learning, ZSL) 的概念. 在零样本学习的场景中, 测试实例所属的类别并没有在训练阶段出现过, 而学习的目的正是对这些没有标注的实例进行识别或分类. 由于在零样本学习中, 训练样本和测试样本对应的标注空间是不同的, 因此, 可将零样本学习视为迁移学习的特例, 属于异质迁移学习 (heterogeneous transfer learning) 的范畴[1]. 随着近年的发展, 零样本学习已经逐渐脱离迁移学习, 成为一个独立的研究方向.

零样本学习范式的提出, 为目前分类任务中广泛存在的训练类别不能覆盖测试类别的实际问题提供了一种解决方案, 也为识别从未见过的数据类别提供了可能. 在零样本学习范式下训练的分类器, 不仅能够识别出训练集中已有的数据类别, 还可以对来自未见过的类别的数据进行推理判断. 这使得计算机具有知识迁移的能力, 避免了训练数据类别需要覆盖所有测试类别的限制, 更加符合人们生产实际的需要. 近年来, 零样本学习已经被广泛地运用到计算机视觉、自然语言处理等多个领域. 并在图片识别[2−4]、视频动作识别[5−8] 和文本翻译[9,10] 等任务中取得了重要进展.

与一些经典的学习范式相比, 零样本学习由于提出的时间较短, 因此相关技术发展的也并不十分成熟, 相关综述文献也较少. 目前, 对零样本学习技术做出了系统阐述, 并具有一定影响力的有 Xian 等人[11]、Fu 等人[12]、Wang 等人[1]、冀中等人[13] 的工作. 其中, Xian 等人[11] 的工作主要聚焦于对一些经典零样本分类模型的概括性总结和评判标准, 并基于提出的标准对一些分类模型进行统一性能测试; 文献 [12] 则对零样本分类任务及其相关领域做了更加全面的介绍, 并对广义零样本分类任务做出了更加全面的讨论. Wang 等人[1] 的工作则首次对零样本分类问题中的不同学习方式进行了正式的定义; 而冀中等人[13] 的工作则按照时间线索讨论了零样本学习的发展历史和技术要点. 整体上而言, 上述文献的共性是侧重于讨论零样本分类的发展现状, 尤其侧重于对图像分类领域的技术讨论. 本文以图和零样本学习的相关性为背景, 讨论了图学习在零样本学习中的应用, 包括但不限于图像分类任务, 旨在让读者了解零样本学习与图学习之间的关联.

本文首先在第 1 节对零样本学习范式和图数据进行相关了阐述, 阐明了零样本学习的发展过程、图数据学习提出的背景, 以及二者的定义和基本相关技术. 并在第 2 节依据图数据学习在零样本学习中不同的利用方式分类着重介绍了零样本图学习方法所涉及到的技术. 第 3 节首先介绍了零样本图学习任务中的评估准则, 以及目前零样本图学习所涉及到的应用场景和数据集, 并分析了目前零样本图学习中典型模型的实验结果. 第 4 节则指出了零样本图学习进一步研究中需要解决的问题以及未来可能的发展方向.

1 概述

零样本图学习是指依据特定类别的数据, 利用辅助信息和先验知识, 并在知识组织利用和模型训练的过程中引入图结构作为辅助, 从而实现对其他类别数据的预测或识别的技术. 这一学习范式目前已经在计算机视觉和自然语言处理领域中得到了广泛的研究.

1.1 零样本学习

1.1.1 零样本学习的定义

零样本学习并不是完全不需要训练样本, 其中的“零样本”是指测试实例对应的类别在训练阶段可以是“零样本”的. 零样本学习范式目的在于研究对于特定

的某些类缺失对应的训练样本情况下, 训练模型在使用其他类的训练样本训练后是否仍然可以对这些特定类的输入做出正确的预测.

因此, 零样本学习问题的解决需要辅助信息的帮助以获得从源标注空间到目标标注空间的知识迁移, 这种辅助信息通常是类别之间的关系. 在零样本学习中, 训练样本所对应的源特征空间, 和测试样本所对应的目标特征空间, 是相同的; 但是训练样本所对应的源标注空间 (又称可见类别, seen class), 和测试样本所对应的目标标注空间 (又称未见类别, unseen class), 则是不同的. 如果目标标注空间与源标注空间存在交集, 这种情况被称为广义开集学习 (generalized open set recognition)[14], 也称广义零样本学习 (generalized zero-shot learning, GZSL), 否则称为狭义零样本学习[1,13], 即目标标注空间和源标注空间完全不同. 在没有特殊提及的情况下, 零样本学习一般指狭义零样本学习.

虽然零样本学习是为了解决图像分类领域中实际类别数量远多于数据集所能提供的类别数量的问题而提出的, 但随着技术的发展, 零样本学习已经不仅只在计算机视觉领域发挥作用, 在自然语言处理领域, 尤其是文本相关的任务中也有了重要应用. 为了统一起见, 以下针对图像领域中的零样本学习做出符号定义. 文本领域内的零样本学习问题与之相比, 缺少视觉空间部分.

定义样本-标签对组成的训练集为$D^{tr}=\left\{(x_i^{tr},y_i^{tr})\in X^{tr}\times Y^{tr}\right\}_{i=1}^{N_{tr}}$, 其中, x_i^{tr}是训练集视觉特征空间X^{tr}中第i个样本的特征, y_i^{tr}是训练集标注空间Y^{tr}中第i个样本的类别标签, 即可见类别标注, N_{tr}是训练样本的数量. 即$X^{tr}=\{x_i^{tr}|i=1,2,\cdots,N_{tr}\}$, $Y^{tr}=\{y_i^{tr}|i=1,2,\cdots,N_{tr}\}$. 类似的, 可以定义样本-标签对组成的测试集$D^{te}=\left\{(x_j^{te},y_j^{te})\in X^{te}\times Y^{te}\right\}_{j=1}^{N_{te}}$, 其中, x_j^{te}是测试集特征空间X^{te}中第j个样本的特征, y_j^{te}是测试集标注空间Y^{te}中第j个样本的类标签, 即未见类别标注, N_{te}是测试样本的数量. 在狭义零样本学习中, 训练样本与测试样本的类标签标注互不相交, 即$Y^{tr}\cap Y^{te}=\emptyset$, 令$Y=Y^{tr}\cup Y^{te}$, Y即总标注空间 (或称语义空间); 而在广义零样本学习中, $Y^{tr}\cap Y^{te}\neq\emptyset$, 且$Y^{tr}\subset Y^{te}$. 令$X=X^{tr}\cap X^{te}$, X是总视觉特征空间.

在语义空间中, 每一个类别或属性对应一个向量表示, 称为类别原型 (class prototype). 定义$T^s=\left\{t_i^s\right\}_{i=1}^{N_s}$为可见类别的类别原型, $T^u=\left\{t_j^u\right\}_{j=1}^{N_u}$为未见类别的类别原型. 令$T=T^s\cup T^u$表示所有类别的语义原型空间. 因此, 在零样本学习中, 至少存在着这样两个任务, 一个是类别标注映射到类别原型的可逆函数$\pi(\cdot):Y\leftrightarrow T$, 另一个是根据样本视觉特征$X$中的特征来获得类别原型$T$的函数$f(\cdot):X\rightarrow T$. 图 1 给出了零样本学习的一般流程.

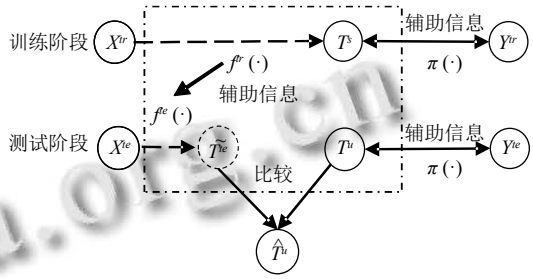


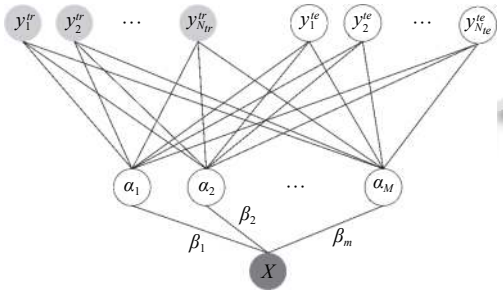
图 1 零样本学习的一般流程 (计算机视觉领域)

零样本学习的基本思想就是利用训练阶段中训练样本X^{tr}和可见类别的类别原型T^s的关系学习到训练阶段的映射$f^{tr}(\cdot)$, 并利用辅助信息将此映射推广到测试阶段的映射$f^{te}(\cdot)$, 再利用相似性比较, 如 K 近邻 (K nearest neighbor, KNN) 度量等, 从而完成对不存在于训练阶段的类别的实例进行推理判别.

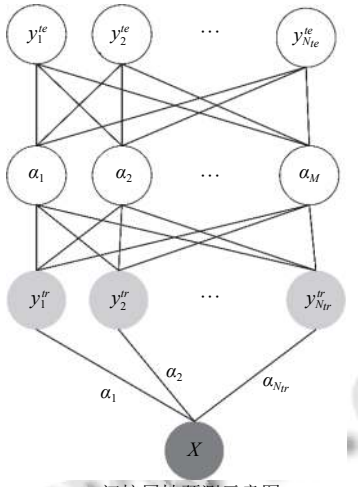
1.1.2 零样本学习的发展历程

2009 年, Lampert 等人提出了一种基于属性的类间迁移学习机制, 即直接属性预测 (DAP) 和间接属性预测 (IAP)[3]. 这两种学习机制在零样本发展过程中有着非常重要的奠基作用, 并持续影响着直到现在的零样本学习方法. 图 2 给出了 DAP 和 IAP 方法的示意图. 在得到图片特征X后, DAP 通过在训练阶段得到的属性预测器α_i预测输入图片所具有的属性, 进而推断输入图片所具有的标签; 而 IAP 首先预测输入图像的类标签, 并根据标签对应的属性指示向量, 间接得到输入图像的属性特征估计. 这一开创性工作利用贝叶斯定理和支持向量机 (support vector machine, SVM), 依据实例包含属性的概率和实例属于类别的概率来预测最终结果. 虽然文中没有提及零样本学习, 但训练集与测试集没有交集, 而且测试集中不包含训练样本所包含的标签集, 这在本质上已经符合了零样本学习的定义. 同年, Palatucci 等人[15] 正式提出了零样本学习的概念. 这项工作以公式化的方法定义了零样本学习问题, 并验证了零样本学习方法的可行性.

在零样本学习技术提出之初[16–20], 主要的研究方法是使用浅层视觉特征, 如尺度不变特征变换 (scale invariant feature transform, SIFT)、图像灰度直方图等, 将事物属性作为语义空间特征, 利用传统的机器学习方法进行判别. 属性作为一种直接描述事物所具有的性质的抽象刻画, 可以容易的完成从可见类别到未见类别的知识转移, 从而进一步实现对未见类别的推断或预测. 这一概念最早体现在 Larochelle 等人[21] 提出的零数据学习 (zero-data learning) 中, 在定义上与后来正式提出的零样本学习本质相同.



(a) 直接属性预测示意图



(b) 间接属性预测示意图

图 2 Lampert 等人提出的基于属性的类间迁移学习机制

但是, 以属性描述事物之间的关系需要耗费巨大的人工成本, 因为属性的描述是由领域专家来定义的, 并且只针对特定数据集. 为此, 人们提出了两种方式来缓解这种成本消耗. 一种方法是通过可见类别的属性建立属性预测映射, 来获得未见类别的属性[2,3,22,23]. 这种方法属于两阶段预测任务, 因而存在中间任务和目标任务的域转移[24]. 例如, DAP[3] 的中间任务是学习属性分类器, IAP[3] 的中间任务是先预测可见类的后验概率, 然后利用每一类的概率来计算图像的属性后验. 后来, 这种两阶段方法已扩展到属性不可用的情况. 另一种方式是使用类别的语义描述, 通过自然语言处理领域的相关技术来描述类[25–28]. 目前最普遍应用的方法是由 Mikolov 等人提出的词向量技术[29], 尤其是基于神经语言模型的方法, 包括 CBOW[29]、skip-gram[29]、GloVe[30] 等. 这类方法能够从大型语料库中自动将单词或者句子生成具有语义信息的向量表征. 在使用类别语义描述的方法中, CONSE[18] 首先预测可见类的概率, 然后通过取前 K 个最可能的可见类的语义的凸结合, 将图像特征投影到语义 Word2Vec[29] 空间, 之后使用 K 近邻方法来得到预测的语义描述.

在零样本学习中, 除了属性预测思想, 另一种思想来源于流形对齐. 由于语义特征和视觉特征是分别提取的, 因此两者对应的空间是相互分离且未对齐的, 但零样本学习需要综合利用两个空间的信息才能够推理出最后的结果, 为此常采用空间映射的方法进行对齐. 例如, Frome 等人[31] 提出 Devise 模型, 使用一种有效的排名损失公式来学习图像和语义空间之间的线性映射. Socher 等人[32] 使用具有两个隐藏层的神经网络来学习从图像特征空间到 Word2Vec[29] 空间的非线性投影. 但是, 由于嵌入空间是一个高维空间, 所以容易出现枢纽化问题 (hubness problem). 该问题是指: 当特征被投影到高维空间中, 一部分测试集的类别可能会成为很多数据点的最近邻, 但其本身所对应的类别之间却不一定具有联系[33]. 在上述的这些方法中, 最终都是使用 K 近邻来获得结果, 因此会受到枢纽化问题的影响. 当视觉特征向语义空间映射时, 会使得空间发生萎缩, 点与点之间更加稠密, 从而加重枢纽化问题[34]. 后续的研究表明, 视觉特征空间作为嵌入空间要比语义空间作为嵌入空间的效果好很多[34,35], 即视觉特征空间比语义空间更具区分性, 因此提出了将语义特征映射到视觉空间的端到端深度嵌入模型[35,36]. 将图像和语义特征嵌入到另一个公共中间空间[37] 是零样本学习方法的另一个方向. 文献 [38] 将视觉特征和语义特征映射到两个独立的潜在空间, 并通过学习另一个双线性兼容函数来测量它们的相似性. 在空间映射的思想基础上, Xian 等人[39]提出了 FGN 模型, 并首次将对抗生成网络 (generative adversarial networks, GAN) 引入零样本学习中, 通过在视觉特征空间中生成具有区分性的

特征数据来完成从原标注空间到目标标注空间的知识转移. 之后, 受到因果关系 (causal ideas) 来研究域适应问题[40] 的影响, Atzmon 等人[41] 将因果推理引入了零样本学习.

图 3 按时间线发展总结了零样本学习发展历程中的重要思想以及相应的模型方法.

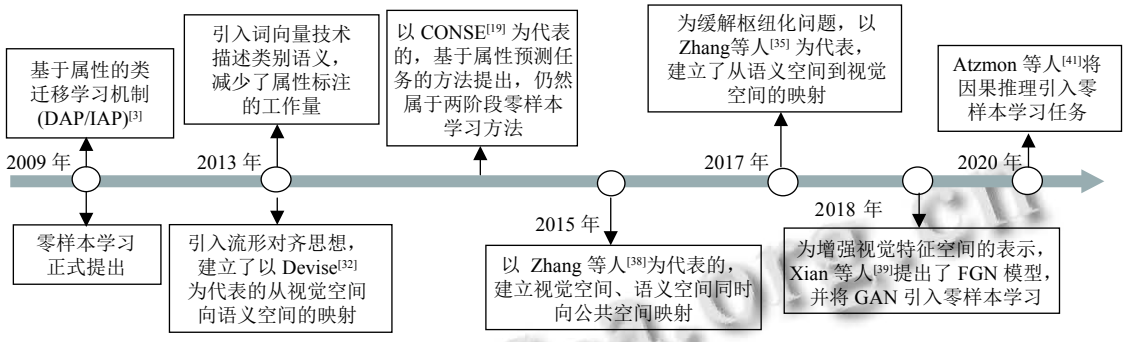


图 3 零样本学习发展历程中的重要思想以及相应的模型方法

1.2 图学习

图学习也称为图数据学习. 图是一种存在于非欧空间的数据结构, 通常由一组节点和边构成. 其中, 边是双向的图称为无向图, 否则称为有向图. 图数据可被用于建模许多真实世界的场景, 具备表达复杂关系的能力, 并被应用在分子化学[42–44]、推荐系统[45–47] 等多个领域.

符号定义如下. 用$G=\{V,E\}$来表示图, 其中, V表示节点集合, $|V|=n$表示图上一共有n个节点; E表示边集合, $|E|=m$表示图上一共有m条边. 通常图用邻接矩阵A来表示, A_{ij}表示节点v_i到节点v_j之间的连接关系, 并且对于无向图来说, $A_{ij}=A_{ji}$. $L=D-A$表示图上的拉普拉斯矩阵, 其中, D是一个$n\times n$的对角阵, D_{ii}表示第i个节点的度且$D_{ii}=\sum_j A_{ij}$. 归一化的拉普拉斯矩阵定义为$L^{\text{norm}}=I_n-D^{-1/2}AD^{-1/2}$, 其中, $I_n\in\mathbb{R}^{n\times n}$是单位矩阵. 又因为归一化拉普拉斯矩阵$L^{\text{norm}}$是一个实对称矩阵, 因此可对$L^{\text{norm}}$做特征分解得到$L^{\text{norm}}=U\Lambda U^{\text{T}}$. 其中, $U=\{u_i\}_{i=1}^n$表示n个相互正交的特征向量, $\Lambda=diag(\{\lambda_i\}_{i=1}^n)$是一个对角阵, λ_i是u_i对应的特征值.

图学习是挖掘图中数据信息和关系信息的算法集合, 通过考虑图的节点特征、邻域节点关联特征以解决实际问题. 图学习一般包括两种方法, 一种是将图转换为表格, 用传统的机器学习方法分析; 另一种是将图建模为网络, 用基于网络的机器学习方法分析. 最近深度神经网络得到快速发展, 相比于传统机器学习方法, 深度神经网络具有更强大的建模能力[48]. 然而, 传统的深度神经网络在全局范围内共享卷积核等参数, 数据需要具有平移不变性, 这是欧式空间数据才具有的特征. 因此, 传统深度神经网络并不能解决图数据学习的需求.

目前的研究成果认为, 图学习方法通常分为 3 大类[49]. 第一, 图嵌入, 也称网络表示学习 (graph/network embedding), 旨在将图 (或图的部分组成) 表示成一个低维向量空间, 同时保留网络 (即对应的图) 的拓扑结构和节点信息, 侧重于学习图的关系结构, 以便在后续的图分析任务中可以直接使用已有的机器学习算法. 第二, 图神经网络正则化 (graph regularized neural networks), 此时图并不直接参与模型训练, 而是充当神经网络的“正则化器”, 从而引导神经网络的损失和数据流向, 并以半监督学习为正则化目标. 第三, 图神经网络 (graph neural networks), 旨在学习具有任意结构的离散拓扑上的可微函数, 并且图节点和边都同时参与模型训练.

图结构具有的点和边自然的可以被理解成事物与事物之间的关系, 这和零样本学习的内涵不谋而合: 零样本学习正是要利用已有样本和与未知样本之间的关系来获得未知样本的表示. 2014 年, Deng 等人[50] 尝试利用类别之间的层次关系构建层次-排除图 (hierarchy and exclusion graphs , HEX graphs), 利用图的边传播信息以获得知识. HEX graphs 虽然不是专门为零样本学习设计的方法, 但是由于图本身包括了类别之间的相互关系, 这种相互关系作为先验知识可以帮助由已知

样本向未见样本进行推导, 从而实现零样本学习. 此后, 越来越多的研究人员开始尝试用图的结构进行零样本推理.

2 零样本图学习方法体系

零样本图学习方法就是利用图学习的相关技术解决零样本问题. 目前, 大多数零样本图学习方法主要针对计算机视觉和自然语言处理等领域的问题. 如第 1.2 节中所述, 图学习方法可以分类为图嵌入、图正则化神经网络、图神经网络 3 种形式[49], 分别代表: (1) 为图结构中的节组件生成低维向量表示; (2) 图充当神经网络的“正则化器”, 从而引导神经网络的损失和数据流向; (3) 以图为载体, 学习具有任意结构的离散拓扑上的可微函数. 本文将这 3 种形式引入零样本图学习领域, 并将其视为在零样本图学习领域中图的利用方法, 从而将零样本图学习体系大致分为 3 类, 即基于知识图谱的零样本学习方法, 零样本图机器学习方法, 和零样本图深度学习方法. 其中, 第 1 类方法主要应用于自然语言处理领域; 后两类方法则在计算机视觉, 尤其是图像分类领域更为常见.

知识图谱是一种通用的揭示实体之间关系的语义网络. 与后两类偏重于使用图的边以权重的形式来度量实体相似度以及信息传播不同, 基于知识图谱的零样本学习方法最大的特征是在图的结构中, 节点和边一般都是具有意义的向量表示, 也因此本类方法多被用于处理文本领域内的任务. 在基于知识图谱的零样本学习中, 一般将不同的知识视为图的节点, 而知识之间的集成或融合形式视为图的边. 在学习过程中, 多采用顺序处理 (如循环神经网络) 或两个实体之间的距离度量 (如翻译式嵌入 (translating embedding, TransE)[51]) 等机器学习方法.

第 2 类是零样本图机器学习方法. 这一类方法的目的是根据已有的图片或视频等视觉材料, 结合相应的语义描述知识, 最后采用传统机器学习方法进行分类、识别任务的演绎推断. 与图正则化神经网络中利用图引导神经网络损失类似, 零样本图机器学习方法旨在建立图以对数据形成约束条件, 并限制传统机器学习方法中的损失函数及信息传递方向. 在图的组织形式中, 一般将视觉材料或语义材料的特征嵌入视为图结构的节点, 节点之间的相似性度量视为图结构的边.

第 3 类是零样本图深度学习方法. 此类方法的目标任务与零样本图机器学习方法相类似, 都是对在训练时期不可见的类别进行分类或识别. 与传统神经网络类似, 图神经网络也可以被认为是一种图特征提取方法, 这种特征提取方法同时考虑了节点本身的特征和节点间的结构信息. 零样本图深度学习方法中图的描述形式与零样本图机器学习类似, 本类方法与前一类方法最大的区别在于其直接在图结构上进行卷积操作.

2.1 基于知识图谱的零样本学习

知识图谱本质上是一种语义网络的形式的知识库, 具有有向图结构. 其中的节点代表实体或概念, 而图的边代表实体/概念之间的各种语义关系, 主要用于描述物理世界中的概念和内在关系. 知识图谱一般使用三元组表示 (head entity, relation, tail entity), 简写为 (h, r, t), 即头实体 (head entity) 和尾实体 (tail entity) 之间的关系 (relation).

知识图谱中边代表的语义关系是节点代表的知识实体间迁移的方式, 这和零样本学习中知识需要联系可见类别和未见类别的内涵是一致的, 即通过某些已经获得的知识的结合来推理出新的知识. 知识图谱的思想目前也被广泛的应用于各种零样本学习场景, 如推荐系统[52,53]、问答系统[54,55] 等.

本节根据知识图谱在零样本学习中的利用形式, 分为知识图谱上的零样本学习和利用知识图谱的零样本学习两种. 前者属于动态知识图谱补全问题的范畴, 而后者利用知识图谱的辅助, 从而更好的完成文本处理领域中的任务.

图 4 给出了基于知识图谱的零样本学习训练时期的一般流程图. 在这项流程中, 输入仅有可见类别的知识表示T以及任务$\hat{T}$所涉及的所有知识之间的图结构 (知识图谱). 在这项任务中, 编码器 ENC 通过输入类别、可见类的表示向量、知识间转移关系从而预测新类别的表示向量, 并利用已经存在于知识图谱中的信息指导损失函数.

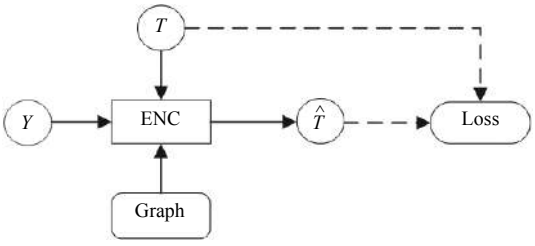


图 4 基于知识图谱的零样本学习方法一般示意图

2.1.1 知识图谱上的零样本学习

一般而言, 大多数的知识图谱是天生不完整的, 因此提出了基于表示学习的知识图谱补全 (knowledge graph complementation, KGC) 算法, 即通过机器学习算法自动地从已有数据中学得新加入知识图谱的节点或边的表示, 从而在较少的人工干预下能自动地根据任务构建特征, 让知识图谱变得更加完整.

根据三元组中的实体和关系是否属于知识图谱中原有的实体和关系, 可以把知识图谱补全分成静态知识图谱补全 (static KGC) 和动态知识图谱补全 (dynamic KGC) 两种. 前者所涉及的实体以及关系均在原始的知识图谱中出现过; 而后者涉及的是不在原始知识图谱中出现的关系以及实体, 从而扩大原有的知识图谱的实体以及关系的集合. 从定义来看, 动态知识图谱补全问题本身即属于零样本图学习的范畴.

在文献 [56] 中, Zhao 等人提出了 JointE 模型, 用来联合学习知识图谱和实体描述嵌入. 在 JointE 中, Zhao 等人根据“实体间通过关系相连, 那么实体是受到关系约束”的这一观点, 提出了基于结构的实体描述, 丰富了节点表达. 由于 JointE 只能用于文本描述的实体, Wang 等人[57] 则提出了当知识图谱应用于非文本知识, 即使用在视觉知识上的实体表示方法—TransAE, 在利用实体和关系之间的结构知识的同时, 保留了实体所具有的知识本身. 与传统方法相比, 多模态知识的引入极大地提高了模型的性能.

上述两种方法虽然都被用于知识图谱上的实体零样本嵌入, 但在训练阶段仍然需要大量标注语料充当监督信息. Li 等人[58] 则针对现有的大多数知识图嵌入模型都是有监督学习范式下的产物, 并且在很大程度上依赖于可获得的标记训练数据的质量和数量这一问题做出了改善. 他们提出了一个两阶段的方法来适应无监督的实体名称嵌入, 随后基于子空间投影的思想, 利用监督模型联合学习子空间中的投影矩阵和知识表示.

事实上, 在大多数知识图中, 通常都有对实体的简明描述, 也就是实体属性. 为了利用这些描述, 从而提高知识图谱嵌入表达的质量, Xie 等人[59] 充分利用实体描述信息提出了一种新的表示学习方法 DKRL (description-embodied knowledge representation learning), 目的在于嵌入实体时同时建模关系, 并在知识图谱补全和实体分类在两个任务上取得了效果. Ding 等人[60] 则提出了使用双向门控递归单元网络 (bidirectional gated recurrent unit network, Bi-GRU) 的方法对实体描述建模, 并建立联合学习实体结构知识和实体描述知识的模型, 加深了知识图谱内外实体之间的有效关联性. 其核心思想是认为相似的实体应该在结构和文本特征空间中具有相似的表示, 即实体的两层结构表示均应具有相似性.

2.1.2 知识图谱辅助的零样本学习

知识图谱作为一种提供实体间显式关系的图结构, 能够天然提供文本处理任务中所需要的辅助信息.

针对从文本中识别属性的任务, Imrattanatrai 等人[61] 针对当难以为每个属性准备训练句的情况下, 利用从知识图结构的不同组件的嵌入获得的属性的表示, 并通过与模型相结合, 使得在没有可用的训练语句的情况下能够识别属性.

针对语义歧义消除任务 (word sense disambiguation, WSD), Kumar 等人[62] 提出了结合意义嵌入的扩展意义嵌入模型 EWISE (extended WSD incorporating sense embeddings), 从有意义的注释数据、字典定义和词汇知识库的组合中学习信息, 通过在连续的语义嵌入空间而不是传统离散的标签空间上进行预测来执行词义嵌入.

2.2 零样本图机器学习

知识图谱作为一种显式地表达知识及它们之间相互联系的图结构, 在计算机视觉领域中也存在着广泛应用. 在这一类应用中, 知识图谱更多的是作为一种先验知识库, 提供实体之间明确的转移关系, 并利用这种知识关联传播信息, 得到未见类别的知识. 由于任务驱动的不同, 模型的输入不仅包括了知识表示, 还包括文本知识以外的其他模态的知识. 虽然同样是利用图的结构关系, 但与基于知识图谱的零样本学习关注知识间的上下义语义关系不同, 计算机视觉的领域的任务在讨论知识间关系属性的同时, 更加关注知识间的相似性关系. 根据图数据的利用形式不同, 可以将面向计算机视觉领域中的零样本图学习方法分为两类, 分别在第 2.2 节和第 2.3 节进行总结.

图机器学习的思想与图正则化神经网络类似, 旨在建立图以对数据形成约束条件, 并进一步地学习预测图的属性. 虽然引入了图数据的表示方式, 但是在数据利用及训练等方面, 仍然遵循传统机器学习算法, 图的主要作用是限制传统机器学习方法中的损失函数及

信息传递方向. 在图机器学习中, 一般针对类别建立图, 即将类别的语义特征或视觉特征作为节点, 特征间相似度作为边[63−65], 也有方法采取节点间的位置关系等非相似度度量的方法作为边[66,67]. 图 5 给出了零样本机器学习训练时期的一般流程图.

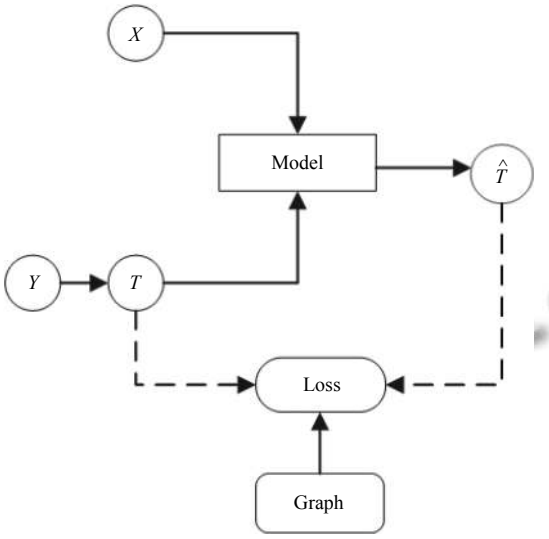


图 5 零样本图机器学习的一般流程图

在图 5 中, X表示视觉特征, T表示由样本标签Y得到的类别语义特征, 零样本图机器学习的目标在于通过输入X和T建立模型 (图中的 Model), 并输出预测的语义特征$\hat{T}$, 而实际的语义特征T作为实际值 (ground truth), 两者相比较得到引导模型训练的 Loss, 创建的图也正是在此时起到作用.

依据图嵌入的不同输出[68], 将零样本图机器学习方法分为 3 类: (1) 基于边嵌入的方法. 这类方法主要侧重于利用特征之间的相似度在图上进行消息传播, 或者对特征进行加强, 从而直接或间接地获得零样本学习的推理结果. (2) 基于节点嵌入的方法. 这种方法的主要思想是利用图上节点和边的信息产生新的节点特征, 并将这种特征应用到解决零样本学习问题的方法中. (3) 基于混合嵌入方法. 混合嵌入指的是对不同类型图组件同时嵌入, 例如同时对节点和边的嵌入. 这类方法主要体现在后续使用子图匹配[69]、图割[70] 等机器学习算法的模型上.

2.2.1 基于边嵌入思想的零样本图机器学习

边嵌入方法利用图的边进行消息传递或利用边的关系以保持相关特征, 是零样本图机器学习方法中应用最为广泛的一种方式. 一般而言, 边描述的是节点特征相似度或节点位置关系.

边嵌入的一种应用方式是消息传递, 即通过边对于节点的连接, 聚合邻居节点信息, 并将其与中心节点自身的信息进行整合. 在利用图的边进行消息传递的方法中, Gao 等人延续了零样本学习问题中属性学习的思想, 提出了一种统一的半监督学习 (semi-supervised learning, SSL) 框架[63], 通过学习嵌入数据点之间关系的最优图来为半监督学习模型生成几何正则化子, 以利用标记图像和未标记图像来学习属性分类器, 最后通过直接属性预测的方式完成零样本分类任务. 而为了缓解一般零样本识别模型仅依赖于未见类别的视觉外观的局限性, 文献 [71,72] 利用视觉特征矩阵度量多目标场景下可见类别与未见类别的相似性, 并融合知识图谱中的语义信息来校准未见类别的预测. 此外, 层次图的利用也是一种比较典型的图消息传递方法, 主要针对类别标签进行建模. 利用层次图进行零样本学习, 类别通常是监督学习中广泛使用的常见类. 在从可见类派生出不可见类的过程中, 每对可见和不可见的类之间的关系通常是从语义空间中的相应原型获得. Deng 等人首次提出了 HEX graphs 的概念[50]. HEX graphs 对标签之间的依赖关系进行显式建模, 将类别原型视作图的节点, 类别之间的联系视作图的边. HEX graphs 的节点采用二值化标签, 利用传统机器学习算法中的条件随机场思想建立分类器. 但是, HEX graphs 是具有确定性或硬约束的概率图形模型, 虽然减少了标签的数量, 带来更精确的推理结果, 但也在一定程度上造成模型拟合不好的问题. 为此, Ding 等人针对标签之间存在的不确定性关系, 将 HEX graphs 中的边替换成“软”联系或概率联系 (soft or probabilistic relations), 由此建立新的层次图模型 pHEX[73], 并在推理过程中, 将 pHEX 模型转换为 Ising 模型[74] 来执行. 类似的, Kordumova 等人将层次图的思想引入了场景分类任务中[75], 为了引导对象和场景之间的知识转移, 研究了类别粒度之间的层次结构, 并针对这些对象建立了层次图. 这种方法可以引导对象在语义嵌入中的表示, 在不使用任何场景图像作为训练数据的情况下识别实例的场景.

边嵌入的另一种应用方式是信息保持, 主要目的是希望在应用机器学习算法时, 数据能够尽量保有原先的特征含义. 从空间映射的角度来说, 由于视觉空间与语义空间存在流形不对齐的问题, Deutsch 等人[76] 基

于多尺度图变换谱图小波 (spectral graph wavelets, SGWs)[77] 对齐算法, 提出了一种基于图上局部多尺度变换的流形对齐框架来解决零样本学习问题. 该方法通过线性投影的方式将语义空间的特征平滑映射到视觉特征空间. 而在文献 [78] 中, Zhong 等人则针对跨模态检索的问题提出了一种跨模态属性哈希模型 (cross-modal attribute hashing, CMAH), 分别对跨模态数据采用图正则化约束以保持各模态的局部结构信息并减少量化损失.

2.2.2 基于节点嵌入思想的零样本图机器学习

节点嵌入的方法是在获得图数据的情况下, 根据已有的图节点特征及特征之间的关系, 生成新的节点特征并用于后续机器学习算法. 节点嵌入的本质是语义空间和视觉空间的对齐问题, 对应的新特征往往是融合了视觉特征和语义特征的共同特点得到的, 因此能够作为连接语义空间和视觉空间的桥梁.

一种典型的节点嵌入方式是对偶图思想. 文献 [79] 中, Long 等人针对视觉特征投影到共享语义空间的单向范式会产生视觉-语义歧义问题, 提出了一种视觉-语义歧义消除的方法 (visual-semantic ambiguity removal, VSAR). 具体是利用对偶图正则化嵌入算法, 同时提取视觉信息和语义信息的共享成分, 并基于两个空间的内在局部结构对齐数据分布, 以减小视觉外观和语义表达之间的差距. 类似的, 文献 [80] 中讨论了给定的语义不足以描述视觉对象的情况, Ding 等人提出了一种基于增强视觉特征和潜在语义表示边缘潜在语义编码器 (marginalized latent semantic encoder, MLSE) 的结构, 利用语义流形中的内在关系, 通过边缘化策略增强视觉语义的泛化能力. MLSE 通过自适应图学习, 实现健壮的图形引导语义编码器, 以此寻找潜在语义表示来更好地描述视觉样本, 有效的缓解了可见类别和未见类别在不同视觉分布上的阻碍等问题.

另一种典型的节点嵌入方法是利用带有权重的二部图 (weighted bipartite graph) 思想[81], 建立视觉空间与语义空间的连接. 这种方法主要是来源于幻影类 (phantom class) 的应用. 幻影类于 2016 年首次被提出[82], 是一种既存在于视觉空间, 也存在于语义空间的一种非真实存在类别, 主要作为连接两种空间的基分类器. 在使用幻影类时, 语义空间和视觉特征空间作为二部图的两个集合, 在每一个空间内, 真实类别 (real class) 和幻影类又分别作为二部图的两部分节点集合. 主要思想是在保留语义关系的前提下, 使幻影类的凸结合尽量靠近真实类别的视觉特征. 幻影类的应用随后在 Chen 等人[83] 的论述中得到了一些改进. 即为了在新的图结构保留一定的邻域结构, 在计算边的权重时, 加入了真实类别邻居的信息, 从而丰富了真实类别和幻影类之间的对应关系.

随机行走是另一种常用的基于节点嵌入方法. 文献 [84] 针对深度模型的选择性学习行为导致视觉特征的区分度降低的问题[85], 受到“分而治之”思想的启发, 提出了一种新颖的、普遍适用的框架—解耦度量学习 (decoupled metric learning, DeML). DeML 是一种基于混合注意力的解耦方法, 通过将嵌入表示解耦到多个注意力特定的学习者, 并以随机行走的方式对像素级对象特征进行加强. 类似的, 文献 [86] 通过限制最大近邻数量和最大后继节点数量, 也以随机行走的方式得到邻域结构信息从而表示节点. 随机行走作为一种无参数的空间注意力方法, 通过在卷积图中深层反应的感受野上进行图传播, 从而能够更全面的对图进行采样, 并进一步地对图节点特征进行增强.

一种与节点嵌入非常相似的方法是基于全图嵌入的零样本学习方法. 基于节点嵌入的方法往往以特征点为融合单位, 但基于全图嵌入的方法应用方式与节点嵌入不同, 以整个流形空间为融合基础. Li 等人[87] 提出了使用矩阵分解策略学习一个视觉对齐的语义图, 在此基础上提出了一种非参数图推理方法, 即流形对齐的图推理 (graph inference with manifold alignment, GIMA). GIMA 不需要学习跨模态视觉语义映射, 而是从不同的模态空间中提取各自的内在流形, 并将它们表示为图结构, 进而通过矩阵分解策略来学习视觉上对齐的语义图, 最后通过简单的图推理算法直接预测新测试图像的类别标签.

2.2.3 基于混合嵌入思想的零样本图机器学习

混合嵌入方法同时针对节点和边进行图操作, 从某种程度上来说能够更多的保留特征信息, 但是也为特征利用带来了一定的困难.

一种混合嵌入的方式是利用最大子图匹配算法. Castanon 等人[66] 提出了一种以用户为中心的方法, 通过创建基于属性和区分关系的稀疏语义图来对查询建模. 同时, 用最大鉴别生成树 (maximally discriminative spanning tree, MDST) 来代替求解时间复杂度为 NP-hard 的精确子图匹配问题 (NP-hard 问题是指无法在多

项式的时间里验证一个解的问题). 该方法通过建模帧图片内的物体以及物体间的位置关系, 通过最大鉴别子图匹配 (maximally discriminative subgraph matching, MDSM) 在线性时间内完成了在没有训练过程的情况下, 直接对视频进行跨模态搜索的目标.

另一种混合嵌入的思想来源于图割 (graph cut) 算法的使用. 文献 [88] 针对文档修复任务, 提出图割的结果可以产生更好的边缘估计. 为此使用完全卷积神经网络 (fully connected neural network, FCNN) 进行语义分割, 通过该阈值概率掩码来构造图, 并利用背景概率图对图割中的边进行剪枝, 从而给出前景和后景的良好估计. 在计算机视觉任务中, 图割法常被用于两阶段零样本学习方法. 文献 [89] 提出分组模拟集成 (grouped simile ensemble, GSE) 框架, 以明喻 (similes) 作为显式属性标注, 以此建模图片之间视觉表达上的相似性. 首先使用图割算法和聚类算法利用视觉相似度从中发现隐含的属性, 并判断属于哪一个明喻簇, 再利用语义相似度判断具体语义类别. 此外, Huang 等人首先提出了一种基于超图的属性预测器 (hypergraph-based attribute predictor, HAP)[90]. HAP 利用超图来刻画数据中属性的高阶和多重关系, 利用类信息和任何可用的辅助信息, 将属性预测问题转化为正则化超图割问题. 在 HAP 超图的设定中, 每个顶点对应于一个样本, 而超边是共享相同属性标签的顶点集.

2.3 零样本图深度学习

目前, 虽然图深度学习方法已经有了广阔的发展, 但在零样本图学习中, 应用主要聚焦于图卷积神经网络. 总体来说, 基于图深度学习的零样本学习方法可以大致分为两类, 分别是图信息基于谱域的传播方法和图信息基于空间域的传播方法. 基于频谱的方法从图信号处理的角度引入滤波器来定义图卷积, 其中图卷积操作被解释为从图信号中去除噪声; 而基于空间的方法将图卷积表示为领域聚合的特征信息, 当图卷积网络的算法在节点层次运行时, 图池化模块可以与图卷积层交错, 将图粗化为高级子结构. 谱域上的图操作具有扎实的理论基础, 根据图谱理论和卷积定理, 将数据由空间域转换到谱域做处理; 而空间域图操作不依靠图谱卷积理论, 直接在空间上定义卷积操作, 具有较强的灵活性. 图 6 给出了零样本图深度学习的一般流程.

零样本图深度学习的一般流程与零样本图机器学习的流程相似, 只是零样本图深度学习直接在图上进行卷积微分操作.

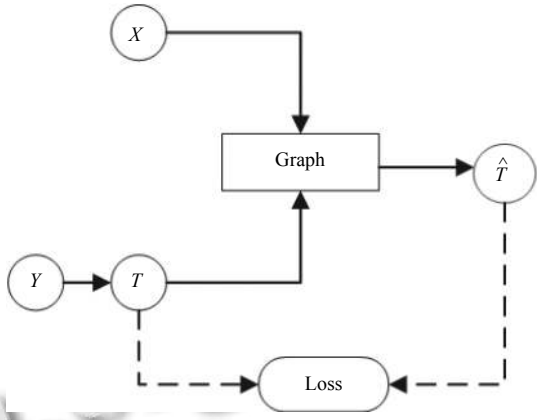


图 6 零样本图深度学习的一般流程

2.3.1 基于谱域的传播方法

谱图卷积建立在图信号处理的基础上, 对图像特征处理具有较大的作用. 尤其在当输入了大量的图像特征、语义特征及图像-语义相关特征时, 这些特征在不经处理的情况下大量使用可能会导致特征利用率降低, 从而使零样本学习的准确率降低. 谱图卷积从信号处理的角度, 缓解了信号嘈杂问题, 提高了特征利用率. 基于谱域的传播方法主要遵循如下公式来更新每层的消息传播:

$$H^{(l+1)}=\delta(D^{-1/2}AD^{-1/2}H^{(l)}W_{\theta})$$

其中, $H^{(l)}$为上一层谱域传播的输出, 将作为本层传播的输入; $H^{(l+1)}$表示本层的传播结果; $\delta(\cdot)$表示激活函数, W_{θ}表示权重矩阵, A是对应图结构的邻接矩阵, D是对应图结构的度矩阵. 在谱域传播中, A一般是一个$\{0,1\}$的二值矩阵, 表示两个节点之间是否有边的存在; 如果边具有权重, 则A根据相应的权重值进行调整.

Shen 等人[67] 首先针对图像-草图检索问题, 建立了零样本草图图像哈希模型 (zero-shot sketch-image hashing, ZSIH). ZSIH 利用克罗内克融合层 (Kronecker fusion layer) 和图形卷积来缓解草图图像的异构性, 以此增强数据之间的语义关系. 在密集图传播模块 (dense graph propagation, DGP) 中[91], 通过增加节点连接以丰富图表达, Kampffmeyer 等人[91] 也提出使用谱域卷积的方式来增强特征表达. 为了利用外部知识信息来显示类别之间的关系, Gao 等人[92] 提出了基于结构化知识图的端到端零样本动作识别框架, 设计了双流图卷积网络 (two-stream graph convolutional network, TS-GCN),

使用谱图卷积减少建模动作-属性、属性-属性和动作-动作之间时产生的信号噪声. 类似地, 在图像注释任务中, 为了缓解多义词带来的信号偏差, 以及语义损失所造成的模型泛化问题, Wang 等人[93] 使用归一化拉普拉斯矩阵的谱域卷积来建立单词向量和图像之间的映射, 使得目标和源标签可以一起训练, 从而缓解了多义词和广义零样本设置中的强偏问题 (指预测结果偏向可见类别的情况). Bucher 等人[94] 提出的 ZS3 网络 (zero-shot semantic segmentation) 可以用来解决零样本图像分割问题, 其模型结合了丰富的文本和图像嵌入, 并包含大量上下文信息, 因此谱域卷积的方式可以尽可能的减少由大量特征输入所带来的信号噪声. 此外, 谱域卷积也被用来平衡损失. Xie 等人[95] 提出一种区域图嵌入网络 (region graph embedding network, RGEN) 来捕捉图像不同区域之间的关系, RGEN 将转移损失和平衡损失纳入框架, 缓解了一般零样本学习模型中的极端区域偏向问题, 降低了图像某些区域特征包含的噪声.

2.3.2 基于空间域的传播方法

基于空间域的图深度学习方法通过信息聚合继承的思想来定义相关图操作, 目的在于利用邻域节点特征增强中心节点的特征表示. 一般而言, 空间域传播方法的特征输入数量或种类比基于谱域的传播方法少. 基于空间域的传播方法主要遵循如下公式来更新每层的消息传播:

$$H^{(l+1)} = \delta(\hat{A}H^{(l)}W_{\theta})$$

其中, $H^{(l)}$为上一层空间域传播的输出, 将作为本层传播的输入; $H^{(l+1)}$表示本层的传播结果; $\delta(\cdot)$表示激活函数, W_{θ}表示权重矩阵, $\hat{A}$是对应图结构的归一化邻接矩阵, 邻接矩阵既可以是一个{0,1}的二值矩阵, 表示两个节点之间是否有边的存在; 也可以是具有边权重的权重矩阵.

Wang 等人[96] 将语义嵌入作为输入, 首次将图网络应用到了图像识别领域, 将零样本学习问题看做一个分类器权重回归问题, 用视觉分类器对应权重作为监督, 建立类别语义间的图知识结构, 显著地提高了零样本图识别的准确率. Yan 等人针对零样本目标检测问题, 提出了基于图卷积网络的语义保持图传播模型 (semantics-preserving graph propagation model, SPGP)[97]. SPGP 结合了一个图构造模块和两个语义保持的图传播模块, 来缓解视觉-语义鸿沟, 同时利用结构知识和描述知识加强了语义表示. 在多标签分配任务中, Lee 等人[98] 利用知识图中定义的不同关系, 最大限度地通过信念传播 (belief propagation, 也称消息传播)[99] 的方式丰富语义空间中传播的标签表示和信息.

在空间域中, 除了一般的图卷积传播方法, 受到传统神经网络中注意力机制、残差模块等思想的启发, 这些网络设计也被应用到空间域图神经网络方法中来. Xiao 等人[100] 提出了一种快速混合模型 ARGCN-DKG (attention based residual graph convolutional network on different types of knowledge graphs) , 通过引入残差机制和注意机制, 整合不同的知识图, 提高不同类别间知识转移的准确性. Zhang 等人[101] 通过图生成模型来显式地建模关系, 其提出的可转移图生成 (transferable graph generation, TGG) 模块旨在捕获类概念、属性和可视化实例之间的关系, 由多头图注意机制引导邻近信息聚合, 从而缓解域转移的适应问题. Wang 等人[102] 提出的注意力图神经网络 (attentive graph neural network, AGNN) 则对帧之间的关系建立图结构, 通过注意力机制有效地捕获了两帧之间的相关性, 同时使用递归消息传递在图上迭代地传播信息, 从而捕获视频帧之间的高阶关系, 并从全局视图获得更优化的结果.

3 评估准则与数据集

目前的零样本图学习主要被应用在计算机视觉和自然语言处理领域. 计算机视觉领域的典型应用包括物体识别[82,90,91,96]、图像检索[67,78]、图像语义分割[94]、视频动作识别[66,92,103] 等. 自然语言处理领域的典型应用如知识图谱表示学习[56,59,60]、知识问答[104] 等. 本节重点从应用场景角度介绍零样本图学习的评估准则, 常用数据集以及目前的最佳效果.

3.1 评估准则

零样本图学习方法的评估准则遵循一般零样本学习的评估方法, 一般而言有如下 4 种:

(1) Top-K 精度. Top-K 精度通常用 Hit@K 来表示, 指的是预测结果中最有可能的前 K 个中包含实际结果的概率. Top-K 精度评估被广泛的应用于自然语言处理和计算机视觉领域. 但一般而言, 自然语言处理中的任务多只报告 Top-1 准确率, 而计算机视觉领域中, 尤其是图像识别任务, 由于类别众多, 通常会报告更大的 K 值精度. 例如在 ImageNet 数据集[105] 上的零

样本识别任务, 研究人员多同时报告 Top-1、Top-2、Top-5、Top-10、Top-20 五种识别精度.

(2) $F1$-score. 即 $F1$ 分数, 是另一种常被使用的评价指标, 尤其是在分类的任务中. 它同时兼顾了分类模型的精确率 (*precision*) 和召回率 (*recall*), 可以看作是模型精确率和召回率的一种调和平均. $F1$ 分数的计算方法为:

$$F1\text{-score} = 2 \times \frac{precision \times recall}{precision + recall}$$

其中, *precision* 指被分类器判定正例中的正样本的比重, *recall* 指被预测为正例的占总的正例的比重.

(3) 类平均准确度 (mean average precision, *mAP*). 一般而言, 大部分任务会采用平均准确率对整体数据集进行评价. 由于在图像领域中实际存在的长尾分布问题, 导致数据集存在大量的样本间数量分布不均衡的情况, 此时如果使用平均准确率进行评价, 则不能较好地反映出大数量样本和小数量样本之间分类准确度的差异. 因此, 目前广泛采用全类平均准确度 *mAP* 作为零样本分类评价指标, 即先对每个类统计类内的分类准确度, 再通过求均值计算类平均准确度, 类平均准确度计算公式为:

$$mAP = \frac{\sum_{n=1}^{N} Acc_{y_i^{te}}}{K}$$

其中, K 表示未见类别的总数, $Acc_{y_i^{te}}$ 表示未见类 i $(1 \leqslant i \leqslant K)$ 中第 i 个类的分类准确度.

(4) 谐调平均准确度 (harmonic mean accuracy, 也称 *H*-value). 目前尚缺乏对广义零样本分类性能的较为统一的评价准则, 相关学者也在积极地探索中. 例如, 为了能够全面衡量算法在训练数据和测试数据上的分类性能, Xian 等人[11] 提出了谐调平均准确度作为广义零样本分类性能评价指标, 谐调平均准确度的计算公式为:

$$H\text{-value} = 2 \times \frac{Acc_{y^{tr}} \times Acc_{y^{te}}}{Acc_{y^{tr}} + Acc_{y^{te}}}$$

其中, $Acc_{y^{tr}}$和$Acc_{y^{te}}$分别表示在可见类别标签和未见类别标签上得到的类平均准确度.

3.2 数据集与目前最高水平 (SOTA)

3.2.1 计算机视觉

在计算机视觉领域的应用中, 零样本图学习目前主要涉及到以下几种应用: 目标检测、图像识别、动作识别.

在零样本目标检测任务中, 有 3 个常用数据集, 分别是: PASCAL VOC 2007+2012[106], ILSVRC 2017[107] 和 MS COCO 2014[108]. PASCAL VOC 数据集主要含有 4 大类别, 分别是人、常见动物、交通车辆、室内家具用品. ILSVRC 是 ImageNet 的一个子集, 因为考虑到目标规模、图像杂乱程度、目标平均实例数等不同因素, ILSVRC 对每个基本类别进行了仔细的选择. MS COCO 则是一个专门为对象检测和语义分割任务而设计的数据集, 由 80 个类别组成. 表 1 给出了上述 3 个数据集的相关描述和目前零样本图学习的 SOTA. 其中评估准则一栏的括号内表示评价指标.

表 1 零样本目标检测常用数据集

数据集	PASCAL VOC 2007+2012	ILSVRC	MS COCO
训练集大小	7 768	456 567	82 783
验证集大小	8 333	20 121	40 504
测试集大小	5 011	—	—
标注对象数量	52 090	534 309	—
评估准则*mAP* (%)	66.4[97]	20.87[97]	35.4*[97]

注: *表示在标准IoU=0.5时取得的结果, MS COCO由于测试集上没有标签注释, 所提出的模型是在MS COCO 2014的训练集中的样本上训练的, 而检测性能是在验证数据集中的图像子集上测量的.

在图像识别任务中, 图像数据集包括动物类别数据集 AwA[3]、AwA2[109], 鸟类数据集 CUB[110], 场景类数据集 SUN Attribute[111]、Places2[112], 混合类别 (包含人物、动物、风景等) 的数据集 aPY[2] 和 ImageNet[105] 等, 其中前 4 个数据集提供属性标注, 而 ImageNet 没有提供属性标注. 此外, CUB 和 SUN 是细粒度图像分类数据集, 其中的图像类间差异较小, 对零样本图像分类的挑战性也较大. 其中属性数据集的相关内容如表 2. 由于 ImageNet 并没有属性标注, 因此在实验中常采取不同其他数据集的组织利用形式. 在实验中, 通常会采用“2-hops”“3-hops”和“All”, 即根据 ImageNet 标签层次结构考虑与原始看到的 ImageNet 1K 类相距 2 跳、3 跳和所有跳的所有类, 对应于 1 549、7 860 和 20 842 类. 目前的 ZSL 和 GZSL 的 SOTA 由 Xiao 等人[100] 给出, 以 Hit@K 作数据集评价标准. 如表 3 所示.

在图像标注任务中, 常用数据集包括 NUSWIDE[113], COCO[108], IAPR TC-12[114] 和 Corel5k[115]. 其中 NUSWIDE 是一个多标签场景数据集, 并可以用于图像文本匹配; IAPR TC-12 包含拍摄于世界各地的静态自然图像, 内含各种静态自然图像的剖面图, 包括各类运动或行动的图像, 可以用于评估自动图像标注方法并研究其对

多媒体信息检索的影响; Corel5k 数据集是图像实验的事实标准数据集, 涵盖多个主题, 并可以用于科学图像实验. 目前的 SOTA 由文献 [93] 给出. 数据集相关由表 4 给出.

表 2 零样本图像分类属性数据集

数据集	CUB	SUN	AWA	aPY
类别数量	200	717	50	32
实例数量	11 788	14 340	30 475	15 339
属性数量	312	102	85	64
标注级别	图片	图片	类别	图片
标注类型 (布尔型或实值)	均有	均有	均有	均有
评估准则*mAP* (%)	76.1[95]	63.8[95]	86.5[65]	63.3[65]

表 3 零样本图像分类数据集 ImageNet 上不同任务的 Hit@1 (%)

方法	2-hops	3-hops	All
ZSL	25.5[100]	6.4[100]	3[100]
GZSL	24.8[100]	6.2[100]	2.9[100]

表 4 零样本图像标注数据集

数据集	NUSWIDE	COCO	IAPR TC-12	Corel5k
图片数量	209 347	122 585	19 627	4 993
标签数量	81	80	291	269
训练集大小	125 449	82 081	17 665	4 493
测试集大小	83 898	40 504	19 627	499
平均标签数量 (每张图片)	2.4	2.9	5.7	3.4
评估准则*F*1-score (%)	18.49[93]	22.13[93]	20.87[93]	22.26[93]

在动作识别任务中, 常用数据集包括 Olympic sports[116], HMDB51[117] 和 UCF101[118]. Olympic sports 是从 YouTube 上下载的, 共 783 段运动员参加 16 种不同运动的视频. HMDB51 来自 YouTube, Google 视频等, 动作类型主要包括: 一般面部动作微笑、面部操作与对象操作、一般的身体动作、与对象交互动作、人体动作. UCF101 在动作方面具有最大的多样性, 动作类别可以分为 5 种类型: 人与物体的互动、仅肢体运动、人与人的互动、演奏乐器、体育. 如表 5 所示.

3.2.2 自然语言处理

自然语言处理中的零样本学习任务主要集中在动态知识图谱补全问题, 以及利用知识图谱进行辅助文本处理的任务, 如语义消歧、实体属性识别.

评估知识图谱嵌入的常用数据集有包括 FB15K[51], FB20K[59]. 一般实验中将 FB15K 中的实体视为 KG 内实体, 将 FB20K 中的额外实体视为 KG 外实体. FB20K 包含 4 组: 头实体和尾实体都在 KG 内 (e-e), 头实体在 KG 外但尾实体在 KG 内 (d-e), 尾实体在 KG 内但头实体在 KG 外 (e-d), 头实体和尾实体都在 KG 外 (d-d). 最优结果由文献 [60] 给出. 经过处理后, 符合训练及测试条件的数据集情况如表 6 所示, 结果由表 7 所示. 更多的数据集使用可以参考文献 [86].

表 5 零样本动作识别数据集

数据集	Olympic sports	HMDB51	UCF101
视频数量	800	6 766	13 320
类别数量	16	51	101
背景形式	动态	动态	动态
评估准则 (%)	59.9±5.3[92]	31.0±3.2[92]	41.6±3.7[92]

注: Gao等人[92]报告了他们的模型在数据集上的准确率±均方差的结果, 但并未指明该结果是否是*mAP*.

表 6 零样本图学习知识图谱嵌入数据集

数据集	实体数量	关系数量	训练集大小	验证集大小	测试集大小
FB15K	14 904	1 341	472 860	48 991	57 803
FB20K	19 923	57 803	18 753	11 586	151

表 7 零样本图学习知识图谱嵌入 SOTA (%)

测试类别	e-e	d-e	e-d
评估准则Hits@10	60.50[60]	34.30[60]	29.60[60]

在语义消歧任务中, 常用数据语料库包括 SensEval-2 (SE2)[119], SensEval-3 (SE3)[120], SemEval-2013 (SE13)[121], SemEval-2015 (SE15)[122]. 其中, SemEval 类的数据集是 SensEval 类数据集的衍生. SE2 的测试集是英语全词任务, 该数据集包含来自华尔街日报的 3 篇文章中的 2 282 个注释. 大多数注释都是名义上的, 还包含动词, 形容词和副词的注释; SE3 是对 SE2 的手工注释的改进. SE13 包括两个消除歧义的任务: 实体链接和词义消歧, 该测试集包含以前版本的统计机器翻译研讨会中的 13 篇文章, 共包含 1 644 个测试实例, 均为名词. SE15 相比于 SE13 更为复杂, 共包括两个领域内 6 087 条测试实例. 目前, 零样本图学习在语义消岐任务上的最优结果均由文献 [62]给出. 如表 8 所示.

表 8 零样本图学习语义消歧数据集 (%)

数据集	SE2	SE3	SE13	SE15
评估准则*F*1-score	73.80[62]	71.10[62]	69.40[62]	74.50[62]

在实体属性识别的任务中, 常用的数据语料库有 NYK10[123] 和 WEB19[61]. 其中, NYK10 中有 54 种属性, 共 99 783 条句子; WEB19 有 271 种属性, 共 45 758 条句子. 目前, 零样本图学习在文本实体属性识别任务

上的最好结果是 Imrattanatrai 等人[61] 的工作结果, 如表 9 所示.

表 9 零样本属性识别任务的评估结果 (%)

数据集	*Accuracy*	*precision*	*recall*	*F*1-score	Hit@1
NYK10	44.1	44.1	68.3	51.1	47.1
WEB19	39.5	41.4	45.0	41.8	42.0

4 未来展望

零样本学习作为机器学习领域中一个新兴的方向, 最近几年取得了飞速的发展. 作为一种衍生于深度学习并且和深度学习有强烈联系的一种学习范式, 零样本学习为难以取得大量学习样本的问题提供了解决方案. 在零样本学习中, 图数据结构的利用使得解决方案能够更好的利用训练样本类别和测样样本类别之间的关系, 从而完成辅助信息的知识迁移. 目前来看, 零样本图学习在未来的研究中存在以下几个潜在的研究方向:

(1) 从数据输入的角度来说, 零样本学习如果想要达到更高的精度, 仍然需要使用属性标注进行学习. 但随着训练类别数量和测试类别数量的增加, 属性标注的工作量也会随之增加. 但现有的网络上已经有许多关于类标签的描述, 因为这些文本内容是非常容易得到的, 可以大大减少零样本学习的工作成本. 因而如何建立大量类别标签之间的知识图谱是一个值得研究的问题.

(2) 从多模态特征融合的角度来说, 目前零样本的一个重要处理方法是基于流形对齐理论, 从语义空间向视觉空间映射, 或者从视觉空间向语义空间映射来获得多模态数据的融合特征. 但是, 这本质上仍然是函数映射的设计, 如果想要提高准确率, 就必须要改进函数映射的设计. 图学习的引入为这种流形对齐的思想提供了一个新的方向, 即同时考虑语义空间与视觉空间的双向映射, 从而更全面的利用数据特征.

(3) 从与其他学习范式结合的角度来说, 单样本学习 (one-hot learning) 是一个与零样本学习十分相似的概念, 指的是在学习过程中, 对于某些类别仅使用一个或少量几个样本, 使模型完成任务. 零样本学习和单样本学习虽然在概念和方法上有一定的相似, 但在具体实现机制上仍有着区别. 如果能将单样本学习的一些思想内涵引入零样本学习中, 例如仅在验证时输入少量测试样本对模型进行精调, 可能会提高零样本学习的结果.

从基础理论角度来说, 零样本学习的任务来源是自然界存在的长尾分布而导致的数据不均衡或难以采集, 其解决问题的思想又来自人类的启发式学习, 与深度神经网络来源于人类大脑神经的连接模型相比, 零样本学习范式缺少实际上的理论支撑. 人类的学习机制一直以来都是心理学界甚至生物界广泛讨论和研究的问题, 一些十分流行的假说例如“图式”模型, 从经验引导的角度对人类学习行为做出了解释. 图式, 指的正是人脑中已有的知识经验的网络, 因此, 零样本学习中, 如果能够引入心理学上人类学习范式的交叉研究, 并构建与人脑中的经验网络相似的知识网络, 或许可以进一步的帮助提高零样本学习的精度.

参考文献

1 Wang W, Zheng VW, Yu H, *et al*. A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 13.

2 Farhadi A, Endres I, Hoiem D, *et al*. Describing objects by their attributes. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 1778–1785.

3 Lampert CH, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 951–958.

4 Li X, Fang M, Chen B. Generalized zero-shot classification via iteratively generating and selecting unseen samples. Signal Processing: Image Communication, 2021, 92: 116115. [doi: 10.1016/j.image.2020.116115]

5 Gan C, Lin M, Yang Y, *et al*. Exploring semantic inter-class relationships (SIR) for zero-shot action recognition. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin: AAAI Press, 2015. 3769–3775.

6 Gan C, Yang Y, Zhu LC, *et al*. Recognizing an action using its name: A knowledge-based approach. International Journal of Computer Vision, 2016, 120(1): 61–77. [doi: 10.1007/s11263-016-0893-6]

7 Liu JG, Kuipers B, Savarese S. Recognizing human actions by attributes. CVPR 2011. Colorado: IEEE, 2011. 3337–3344.

8 Su Y, Xing M, An SM, *et al*. VDARN: Video disentangling attentive relation network for few-shot and zero-shot action recognition. Ad Hoc Networks, 2021, 113: 102380. [doi:

10.1016/j.adhoc.2020.102380]

9 Firat O, Sankaran B, Al-Onaizan Y, *et al*. Zero-resource translation with multi-lingual neural machine translation. arXiv: 1606.04164, 2016.

10 Johnson M, Schuster M, Le QV, *et al*. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 2017, 5: 339–351. [doi: 10.1162/tacl_a_00065]

11 Xian YQ, Schiele B, Akata Z. Zero-shot learning—The good, the bad and the ugly. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3077–3086.

12 Fu YW, Xiang T, Jiang YG, *et al*. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. IEEE Signal Processing Magazine, 2018, 35(1): 112–125. [doi: 10.1109/MSP.2017.2763441]

13 冀中, 汪浩然, 于云龙, 等. 零样本图像分类综述: 十年进展. 中国科学: 信息科学, 2019, 49(10): 1299–1320.

14 Fu YW, Sigal L. Semi-supervised vocabulary-informed Learning. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 5337–5346. [doi: 10.1109/Cvpr.2016.576]

15 Palatucci M, Pomerleau D, Hinton GE, *et al*. Zero-shot learning with semantic output codes. Proceedings of the 22nd International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2009. 1410–1418.

16 Lampert CH, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 453–465. [doi: 10.1109/Tpami.2013.140]

17 Al-Halah Z, Tapaswi M, Stiefelhagen R. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 5975–5984. [doi: 10.1109/Cvpr.2016.643]

18 Norouzi M, Mikolov T, Bengio S, *et al*. Zero-shot learning by convex combination of semantic embeddings. arXiv: 1312.5650, 2013.

19 Jayaraman D, Grauman K. Zero-shot recognition with unreliable attributes. Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014. 3464–3472.

20 Kankuekul P, Kawewong A, Tangruamsub S, *et al*. Online incremental attribute-based zero-shot learning. 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3657–3664.

21 Larochelle H, Erhan D, Bengio Y. Zero-data learning of new tasks. Proceedings of the 23rd National Conference on Artificial Intelligence. Chicago: AAAI Press, 2008. 646–651.

22 Parikh D, Grauman K. Relative attributes. 2011 International Conference on Computer Vision. Barcelona: IEEE, 2011. 503–510. [doi: 10.1109/iccv.2011.6126281]

23 Yu FX, Cao LL, Feris RS, *et al*. Designing category-level attributes for discriminative visual recognition. 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 771–778. [doi: 10.1109/Cvpr.2013.105]

24 Fu YW, Hospedales TM, Xiang T, *et al*. Transductive multi-view zero-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(11): 2332–2345. [doi: 10.1109/TPAMI.2015.2408354]

25 Akata Z, Reed S, Walter D, *et al*. Evaluation of output embeddings for fine-grained image classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 2927–2936.

26 Ba JL, Swersky K, Fidler S, *et al*. Predicting deep zero-shot convolutional neural networks using textual descriptions. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 4247–4255. [doi: 10.1109/Iccv.2015.483]

27 Elhoseiny M, Saleh B, Elgammal A. Write a classifier: Zero-shot learning using purely textual descriptions. Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney: IEEE, 2013. 2584–2591.

28 Qiao RZ, Liu LQ, Shen CH, *et al*. Less is more: Zero-shot learning from online textual documents with noise suppression. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 2249–2257. [doi: 10.1109/Cvpr.2016.247]

29 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2013. 3111–3119.

30 Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language

Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1532–1543.

31 Frome A, Corrado GS, Shlens J, *et al*. DeViSE: A deep visual-semantic embedding model. Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2013. 2121–2129.

32 Socher R, Ganjoo M, Manning CD, *et al*. Zero-shot learning through cross-modal transfer. Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2013. 935–943.

33 Lazaridou A, Dinu G, Baroni M. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing: Association for Computational Linguistics, 2015. 270–280.

34 Shigeto Y, Suzuki I, Hara K, *et al*. Ridge regression, hubness, and zero-shot learning. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Porto: Springer, 2015. 135–151.

35 Zhang L, Xiang T, Gong SG. Learning a deep embedding model for zero-shot learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3010–3019. [doi: 10.1109/Cvpr.2017.321]

36 Changpinyo S, Chao WL, Sha F. Predicting visual exemplars of unseen classes for zero-shot learning. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 3496–3505. [doi: 10.1109/Iccv.2017.376]

37 Zhang ZM, Saligrama V. Zero-shot learning via semantic similarity embedding. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 4166–4174. [doi: 10.1109/Iccv.2015.474]

38 Zhang ZM, Saligrama V. Zero-shot learning via joint latent similarity embedding. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 6034–6042.

39 Xian YQ, Lorenz T, Schiele B, *et al*. Feature generating networks for zero-shot learning. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5542–5551.

40 Gong MM, Zhang K, Liu TL, *et al*. Domain adaptation with conditional transferable components. Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York: JMLR, 2016. 2839–2848.

41 Atzmon Y, Kreuk F, Shalit U, *et al*. A causal view of compositional zero-shot recognition. arXiv: 2006.14610, 2020.

42 Gilmer J, Schoenholz SS, Riley PF, *et al*. Neural message passing for quantum chemistry. arXiv: 1704.01212, 2017.

43 Li YJ, Vinyals O, Dyer C, *et al*. Learning deep generative models of graphs. arXiv: 1803.03324, 2018.

44 De Cao N, Kipf T. MolGAN: An implicit generative model for small molecular graphs. arXiv: 1805.11973, 2018.

45 Van Den Berg R, Kipf TN, Welling M. Graph convolutional matrix completion. arXiv: 1706.02263, 2017.

46 Ying R, He RN, Chen KF, *et al*. Graph convolutional neural networks for web-scale recommender systems. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018. 974–983.

47 Monti F, Bronstein MM, Bresson X. Geometric matrix completion with recurrent multi-graph neural networks. arXiv: 1704.06803, 2017.

48 Sze V, Chen YH, Yang TJ, *et al*. Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 2017, 105(12): 2295–2329. [doi: 10.1109/JPROC.2017.2761740]

49 Chami I, Abu-El-Haija S, Perozzi B, *et al*. Machine learning on graphs: A model and comprehensive taxonomy. arXiv: 2005.03675, 2020.

50 Deng J, Ding J, Jia YQ, *et al*. Large-scale object classification using label relation graphs. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 48–64.

51 Bordes A, Usunier N, Garcia-Durán A, *et al*. Translating embeddings for modeling multi-relational data. Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2013. 2787–2795.

52 Wang HW, Zhang FZ, Xie X, *et al*. DKN: Deep knowledge-aware network for news recommendation. Proceedings of the 2018 World Wide Web Conference. Lyon: International World Wide Web Conferences Steering Committee, 2018. 1835–1844. [doi: 10.1145/3178876.3186175]

53 Wang HW, Zhang FZ, Wang JL, *et al*. RippleNet: Propagating user preferences on the knowledge graph for recommender systems. Proceedings of the 27th ACM

International Conference on Information and Knowledge Management. Virtual Event: ACM, 2018. 417–426. [doi: 10.1145/3269206.3271739]

54 Yoon S, Dernoncourt F, Kim DS, *et al*. A compare-aggregate model with latent clustering for answer selection. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Virtual Event: ACM, 2019. 2093–2096. [doi: 10.1145/3357384.3358148]

55 Ma XY, Zhu QL, Zhou YL, *et al*. Improving question generation with sentence-level semantic matching and answer position inferring. Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020. 8464–8471.

56 Zhao Y, Gao S, Gallinari P, *et al*. Zero-shot embedding for unseen entities in knowledge graph. IEICE Transactions on Information and Systems, 2017, 100(7): 1440–1447. [doi: 10.1587/transinf.2016EDP7446]

57 Wang ZK, Li LJ, Li QD, *et al*. Multimodal data enhanced representation learning for knowledge graphs. 2019 International Joint Conference on Neural Networks (IJCNN). Budapest: IEEE, 2019. 1–8.

58 Li CH, Xian XF, Ai XS, *et al*. Representation learning of knowledge graphs with embedding subspaces. Scientific Programming, 2020, 2020: 4741963. [doi: 10.1155/2020/4741963]

59 Xie RB, Liu ZY, Jia J, *et al*. Representation learning of knowledge graphs with entity descriptions. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press, 2016. 2659–2665.

60 Ding JH, Ma SH, Jia WJ, *et al*. Jointly modeling structural and textual representation for knowledge graph completion in zero-shot scenario. Proceedings of the 2nd Asia-Pacific Web (APWeb) and Web-age Information Management (WAIM) Joint International Conference on Web and Big Data. Macao: Springer, 2018. 369–384. [doi: 10.1007/978-3-319-96890-2_31]

61 Imrattanatrai W, Kato MP, Yoshikawa M. Identifying entity properties from text with zero-shot learning. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Paris: ACM, 2019. 195–204. [doi: 10.1145/3331184.3331220]

62 Kumar S, Jat S, Saxena K, *et al*. Zero-shot word sense disambiguation using sense definition embeddings. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 5670–5681.

63 Gao LL, Song JK, Shao JM, *et al*. Zero-shot image categorization by image correlation exploration. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. Lisboa: ACM, 2015. 487–490. [doi: 10.1145/2671188.2749309]

64 Li AX, Lu ZW, Wang LW, *et al*. Zero-shot scene classification for high spatial resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 4157–4167. [doi: 10.1109/Tgrs.2017.2689071]

65 Fu ZY, Xiang T, Kodirov E, *et al*. Zero-shot learning on semantic class prototype graph. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(8): 2009–2022. [doi: 10.1109/Tpami.2017.2737007]

66 Castanon G, Chen YT, Zhang ZM, *et al*. Efficient activity retrieval through semantic graph queries. Proceedings of the 23rd ACM International Conference on Multimedia. Lisboa: ACM, 2015. 391–400. [doi: 10.1145/2733373.2806229]

67 Shen YM, Liu L, Shen FM, *et al*. Zero-shot sketch-image hashing. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3598–3607. [doi: 10.1109/Cvpr.2018.00379]

68 Cai HY, Zheng VW, Chang KCC. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1616–1637. [doi: 10.1109/Tkde.2018.2807452]

69 Ullmann JR. An algorithm for subgraph isomorphism. Journal of the ACM, 1976, 23(1): 31–42. [doi: 10.1145/321921.321925]

70 Boykov YY, Jolly MP. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. Vancouver: IEEE, 2001. 105–112. [doi: 10.1109/iccv.2001.937505]

71 Chang DS, Cho GH, Choi YS. Similarity-based calibration method for zero-shot recognition in multi-object scenes. Proceedings of the 35th Annual ACM Symposium on Applied Computing. Brno: ACM, 2020. 1096–1103. [doi: 10.1145/3341105.3373931]

72 Chang DS, Cho GH, Choi YS. Zero-shot recognition enhancement by distance-weighted contextual inference. Applied Sciences, 2020, 10(20): 7234. [doi: 10.3390/app10207234]

73 Ding N, Deng J, Murphy KP, *et al*. Probabilistic label relation graphs with Ising models. 2015 IEEE International

Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1161–1169. [doi: 10.1109/Iccv.2015.138]

74 Kadowaki T, Nishimori H. Quantum annealing in the transverse Ising model. Physical Review E, 1998, 58(5): 5355–5363. [doi: 10.1103/PhysRevE.58.5355]

75 Kordumova S, Mensink T, Snoek CGM. Pooling objects for recognizing scenes without examples. Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. Lisboa: ACM, 2016. 143–150. [doi: 10.1145/2911996.2912007]

76 Deutsch S, Kolouri S, Kim K, *et al*. Zero shot learning via multi-scale manifold regularization. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5292–5299. [doi: 10.1109/Cvpr.2017.562]

77 Hammond DK, Vandergheynst P, Gribonval R. Wavelets on graphs via spectral graph theory. Applied and Computational Harmonic Analysis, 2011, 30(2): 129–150. [doi: 10.1016/j.acha.2010.04.005]

78 Zhong FM, Chen ZK, Min GY. An exploration of cross-modal retrieval for unseen concepts. Proceedings of the 24th International Conference on Database Systems for Advanced Applications. Chiang Mai: Springer, 2019. 20–35. [doi: 10.1007/978-3-030-18579-4_2]

79 Long Y, Guan Y, Shao L. Generic compact representation through visual-semantic ambiguity removal. Pattern Recognition Letters, 2019, 117: 186–192. [doi: 10.1016/j.patrec.2018.04.024]

80 Ding ZM, Liu HF. Marginalized latent semantic encoder for zero-shot learning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 6184–6192. [doi: 10.1109/Cvpr.2019.00635]

81 Kőnig D. Gráfok és mátrixok. Matematikai és Fizikai Lapok, 1931, 38: 116–119.

82 Changpinyo S, Chao WL, Gong BQ, *et al*. Synthesized classifiers for zero-shot learning. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 5327–5336. [doi: 10.1109/Cvpr.2016.575]

83 Chen Y, Xiong YH, Gao X, *et al*. Structurally constrained correlation transfer for zero-shot learning. 2018 IEEE Visual Communications and Image Processing (VCIP). Taichung: IEEE, 2018. 1–4.

84 Chen BH, Deng WH. Hybrid-attention based decoupled metric learning for zero-shot image retrieval. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 2745–2754. [doi: 10.1109/Cvpr.2019.00286]

85 Chen BH, Deng WH. Energy confused adversarial metric learning for zero-shot image retrieval and clustering. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019. 8134–8141.

86 Chen LH, Qu YR, Wang ZH, *et al*. Sampled in pairs and driven by text: A new graph embedding framework. The World Wide Web Conference. San Francisco: ACM, 2019. 2644–2651. [doi: 10.1145/3308558.3313520]

87 Li YN, Hu HH, Wang DH. Learning visually aligned semantic graph for cross-modal manifold matching. 2019 IEEE International Conference on Image Processing (ICIP). Taipei: IEEE, 2019. 3412–3416.

88 Ayyalasomayajula KR, Brun A. Historical document binarization combining semantic labeling and graph cuts. Proceedings of the 20th Scandinavian Conference on Image Analysis. Troms: Springer, 2017. 386–396. [doi: 10.1007/978-3-319-59126-1_32]

89 Long Y, Shao L. Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa: IEEE, 2017. 907–915. [doi: 10.1109/Wacv.2017.106]

90 Huang S, Elhoseiny M, Elgammal A, *et al*. Learning hypergraph-regularized attribute predictors. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 409–417.

91 Kampffmeyer M, Chen YB, Liang XD, *et al*. Rethinking knowledge graph propagation for zero-shot learning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 11479–11488. [doi: 10.1109/Cvpr.2019.01175]

92 Gao JY, Zhang TZ, Xu CS. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019. 8303–8311.

93 Wang FX, Liu J, Zhang SW, *et al*. Inductive zero-shot image annotation via embedding graph. IEEE Access, 2019, 7: 107816–107830. [doi: 10.1109/Access.2019.2925383]

94 Bucher M, Vu TH, Cord M, *et al*. Zero-shot semantic segmentation. Advances in Neural Information Processing Systems, 2019, 32: 468–479.

95 Xie GS, Liu L, Zhu F, *et al*. Region graph embedding

network for zero-shot learning. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 562–580.

96 Wang XL, Ye YF, Gupta A. Zero-shot recognition via semantic embeddings and knowledge graphs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6857–6866. [doi: 10.1109/Cvpr.2018.00717]

97 Yan CX, Zheng QH, Chang XJ, *et al*. Semantics-preserving graph propagation for zero-shot object detection. IEEE Transactions on Image Processing, 2020, 29: 8163–8176. [doi: 10.1109/Tip.2020.3011807]

98 Lee CW, Fang W, Yeh CK, *et al*. Multi-label zero-shot learning with structured knowledge graphs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1576–1585. [doi: 10.1109/Cvpr.2018.00170]

99 Yedidia JS, Freeman WT, Weiss Y. Understanding belief propagation and its generalizations. Exploring Artificial Intelligence in the New Millennium. San Francisco: Morgan Kaufmann Publishers Inc., 2003. 239–269.

100 Xiao B, Du YJ, Wu QMJ, *et al*. A fast hybrid model for large-scale zero-shot image recognition based on knowledge graphs. IEEE Access, 2019, 7: 119309–119318. [doi: 10.1109/Access.2019.2935175]

101 Zhang CR, Lyu XQ, Tang Z. TGG: Transferable graph generation for zero-shot and few-shot learning. Proceedings of the 27th ACM International Conference on Multimedia. Lisboa: ACM, 2019. 1641–1649. [doi: 10.1145/3343031.3351000]

102 Wang WG, Lu XK, Shen JB, *et al*. Zero-shot video object segmentation via attentive graph neural networks. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 9235–9244. [doi: 10.1109/Iccv.2019.00933]

103 Zhang CY, Tian YL, Guo XJ, *et al*. DAAL: Deep activation-based attribute learning for action recognition in depth videos. Computer Vision and Image Understanding, 2018, 167: 37–49. [doi: 10.1016/j.cviu.2017.11.008]

104 Bapna A, Tür G, Hakkani-Tür D, *et al*. Towards zero-shot frame semantic parsing for domain scaling. 18th Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017. 2476–2480. [doi: 10.21437/Interspeech.2017-518]

105 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/cvpr.2009.5206848]

106 Everingham M, van Gool L, Williams CKI, *et al*. The PASCAL visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

107 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

108 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

109 Lazaridou A, Bruni E, Baroni M. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014. 1403–1414.

110 Wah C, Branson S, Welinder P, *et al*. The Caltech-UCSD birds-200-2011 dataset. 2011.

111 Patterson G, Hays J. SUN attribute database: Discovering, annotating, and recognizing scene attributes. 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 2751–2758.

112 Zhou B, Khosla A, Lapedriza A, *et al*. Places2: A large-scale database for scene understanding. http://places2. csail. mit. edu. (2017-06-21)[2021-07-22].

113 Chua TS, Tang JH, Hong RC, *et al*. NUS-WIDE: A real-world web image database from National University of Singapore. Proceedings of the ACM International Conference on Image and Video Retrieval. Santorini: ACM, 2009. 48.

114 Grubinger M, Clough P, Müller H, *et al*. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. International Workshop ontoImage. Genoa: OntoImage 2006, 2006. 13–23.

115 Duygulu P, Barnard K, de Freitas JFG, *et al*. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. Proceedings of the 7th European Conference on Computer Vision. Copenhagen: Springer, 2002. 97–112.

116 Niebles JC, Chen CW, Li FF. Modeling temporal structure of decomposable motion segments for activity

classification. Proceedings of the 11th European Conference on Computer Vision. Heraklion: Springer, 2010. 392–405.

117 Kuehne H, Jhuang H, Garrote E, *et al*. HMDB: A large video database for human motion recognition. 2011 International Conference on Computer Vision. Barcelona: IEEE, 2011. 2556–2563.

118 Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012.

119 Palmer M, Fellbaum C, Cotton S, *et al*. English tasks: All-words and verb lexical sample. Proceedings of SENSEVAL-2 2nd International Workshop on Evaluating Word Sense Disambiguation Systems. Toulouse: Association for Computational Linguistics, 2001. 21–24.

120 Snyder B, Palmer M. The English all-words task. Proceedings of SENSEVAL-3, the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona: Association for Computational Linguistics, 2004. 41–43.

121 Navigli R, Jurgens D, Vannella D. Semeval-2013 task 12: Multilingual word sense disambiguation. 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013). Atlanta: Association for Computational Linguistics, 2013. 222–231.

122 Moro A, Navigli R. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver: Association for Computational Linguistics, 2015. 288–297.

123 Riedel S, Yao LM, McCallum A. Modeling relations and their mentions without labeled text. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Barcelona: Springer, 2010. 148–163.