软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(2):370−405 [doi: 10.13328/j.cnki.jos.006146]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于知识的零样本视觉识别综述*

冯耀功[1,2], 于 剑[1,2], 桑基韬[1,2], 杨朋波[1,2]

[1](北京交通大学 计算机与信息技术学院,北京 100044)

[2](北京交通大学 人工智能研究院,北京 100044)

通讯作者: 于剑, E-mail: jianyu@bjtu.edu.cn

摘 要: 零样本学习旨在通过运用已学到的已知类知识去认知未知类.近年来,“数据+知识驱动”已经成为当下的新潮流,而在计算机视觉领域内的零样本任务中,“知识”本身却缺乏统一明确的定义.针对这种情况,尝试从知识的角度出发,梳理了本领域内“知识”这一概念所覆盖的范畴,共划分为初级知识、抽象知识以及外部知识.基于前面对知识的定义和划分,梳理了当前的零样本学习(主要是图像分类任务的模型)工作,分为基于初级知识的零样本模型、基于抽象知识的零样本模型以及引入外部知识的零样本模型.还对领域内存在的域偏移和枢纽点问题进行了阐述,并基于问题对现有工作进行了总结归纳.最后总结了目前常用的图像分类任务的数据集和知识库、图像分类实验评估标准以及代表性的模型实验结果,并对未来的工作进行了展望.

关键词: 零样本学习;初级知识;抽象知识;外部知识;图像分类

中图法分类号: TP18

中文引用格式: 冯耀功,于剑,桑基韬,杨朋波.基于知识的零样本视觉识别综述.软件学报,2021,32(2):370−405. http://www.jos.org.cn/1000-9825/6146.htm

英文引用格式: Feng YG, Yu J, Sang JT, Yang PB. Survey on knowledge-based zero-shot visual recognition. Ruan Jian Xue Bao/ Journal of Software, 2021,32(2):370−405 (in Chinese). http://www.jos.org.cn/1000-9825/6146.htm

Survey on Knowledge-based Zero-shot Visual Recognition

FENG Yao-Gong[1,2], YU Jian[1,2], SANG Ji-Tao[1,2], YANG Peng-Bo[1,2]

[1](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

[2](Institute of artificial intelligence, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Zero-shot learning aims to recognize the unseen classes by using the knowledge of the seen classes that has been learned. In recent years, ‘knowledge+data driven’ has become a new trend but lacking of unified definition of ‘knowledge’ in the current zero-shot tasks of computer vision. This study tries to define the ‘knowledge’ in this field and divided it into three categories, which are primary knowledge, abstract knowledge, and external knowledge. In addition, based on the definition and classification of knowledge, the current works on zero-shot learning (mainly in image classification task) are sorted out, they are divided into zero-shot models based on primary knowledge, zero-shot models based on abstract knowledge, and zero-shot models based on external knowledge. This study also introduces the problems which are domain shift and hubness in this field, and further summarizes existing works based on the problems. Finally, the paper summarizes the datasets and knowledge bases that commonly used in image classification tasks, the evaluation criteria of image classification experiment and the experimental results of representative models. The future works are also summarized and prospected.

Key words: zero-shot learning; primary knowledge; abstract knowledge; external knowledge; image classification

* 基金项目: 国家重点研发计划(2017YFC1703506); 国家自然科学基金(61632004, 61832002, 61672518); 中央高校基本科研业务费专项资金(2020YJS030, 2018JBZ006, 2019JBZ110)

Foundation item: National Key Research and Development Program of China (2017YFC1703506); National Natural Science Foundation of China (61632004, 61832002, 61672518); Fundamental Research Funds for the Central Universities (2020YJS030, 2018JBZ006, 2019JBZ110)

收稿时间: 2020-07-03; 修改时间: 2020-08-11; 采用时间: 2020-09-07; jos 在线出版时间: 2020-10-12

大数据时代的到来,使得深度学习的热度持续高涨.但是同时,深度学习的模型也暴露出了很大的问题,其在训练过程中,特别依赖大规模、强标记的数据,往往意味着要耗费极大的人力和物力.针对此问题,有研究者提出了新的学习方式,比如依靠弱标记数据训练模型的弱监督学习[1];除此之外,如何通过之前已获得的知识并用于新的学习任务或者过程,大大减轻数据标注的压力,也成为人们解决这一问题的新出发点.受这一思想的驱动,元学习(meta learning)[2]、零样本学习(zero-shot learning,简称 ZSL)[3−5]和小样本学习(few-shot learning,简称 FSL)[6]等概念变得火热,并且都有所进展.

人们往往具备知识迁移的能力;在大多数情况下,人们可以结合已有知识,并通过针对未知事物较为具体的文字描述,借助标签-视觉对应关系来认知这一新事物.例如,通过已有知识“斑马拥有大熊猫的黑白色彩、老虎的花纹、马的形态”这样的描述,该人即使从未见过斑马,也能够大致推断出斑马的样子;这即是零样本学习最直观的解释.零样本学习最早见于文献[7],从数学角度来定义,其具体是指:给定一些有标记的训练样本,包含了这些有标记样本的类别,称为可见类(seen classes),也称为源域(source domain),其中,可见类数据记为 X_{tr},标签记为 Y_{tr};同时还有一些无标记的样本,包含这些无标记样本的类,称为不可见类(unseen classes),也称为目标域(target domain),数据记为 X_{te},标签记为 Y_{te}.需要指出的是,在零样本学习中,$Y_{tr}\cap Y_{te}=\varnothing$,模型在源域中训练,在目标域中测试.但是,原始的零样本学习往往又有两种基础扩展:(1) 依据在训练过程中是否使用了无标记的 X_{te},零样本学习又分为归纳式零样本学习(inductive zero-shot learning)(使用了 X_{te})和直推式零样本学习(transductive zero-shot learning)(未使用 X_{te});(2) 依据在测试过程中测试类别是否包含可见类,零样本学习又分为传统零样本学习(zero- shot learning)(不包含可见类)和泛化零样本学习(general zero-shot learning)(包含可见类).如果在训练过程中,有少量的不可见类有标记数据参与到训练过程中,那么零样本学习就转化为小样本学习;零样本学习和小样本学习两者又合称少样本学习(low-shot learning)[8].因此,零样本学习是小样本学习的一种更加极端的情况.

通过上面的定义和所举例子表明,零样本学习本质上是用迁移学习(transfer learning)[9]的思想来解决问题的.即从一个域/任务/分布中学出一个有效的模型,然后迁移到新的域/任务/分布.这与迁移学习中域适应的思想内核是类似的.所以研究者也经常将域适应(domain adaptation)任务与零样本学习任务作比较[3,10,11](文献[12]提到了域不变性).而这两者最大的区别就在于:在零样本问题中,可见类与不可见类的交集为空集.因此,零样本学习可以被视作是一个特殊的迁移学习任务.并由于这种迁移性和随之带来的空间变换,在本领域长久以来存在着域偏移(domain shift)[13,14]问题和枢纽点(hubness)[15]问题:前者是因为数据特征空间的不同所带来的特征表示内容发生偏移的问题;后者是由于数据特征维度的变化所带来的“某些不相关点会成为大多数点的最近邻点”,进而会影响对不可见类数据进行认知的问题.详细内容我们将会结合模型细节在第 4 节展开阐述.

文献[3]中指出:“零样本认知的关键就在于如何挖掘出不可见类与可见类之间语义上相关联的知识”.所以从知识迁移的角度出发,零样本学习的成立也存在一个先验条件,那就是可见类和不可见类数据之间须存在某种形式的知识关联.直观理解,这种关联性越强,模型效果越好.综上,零样本问题中知识的获取和迁移就是其关键核心.基于这样的认识,本文首先对“知识”的获取进行了全面的发掘,将其概念范畴定义了 3 个层次:1. 初级知识,即从数据集中直接可以获取的知识,比如属性、类别标签、视觉特征等;2. 抽象知识,即多个数据包含的知识,比如数据概率、流形分布等;3. 外部知识,比如人类已经建立的知识库等.具体内容我们将在第 1 节详细论述.依据划分好的知识层次,本文进一步梳理了现有的零样本学习工作,即知识的迁移方式,具体论述将放在第 2 节.

本文工作相较于文献[5]中归纳式和直推式的模型分类方法,其基于知识迁移的分类方法能够帮助人们更好地理解当前“数据+知识驱动”的思想潮流,也更加接近零样本学习任务的本质问题,即“在目标域和源域之间,通过何种形式的知识,才能更好地搭建起两者之间的桥梁”.在 Rohrbach 等人[16]的工作中,将零样本学习领域基于知识的迁移方式分为 3 个方向:(1) 利用类别来构建的“从一般到特殊”的层次化的体系结构(例如 WordNet 知识库);(2) 基于提取类别之间通用的可视属性,将不同类的特征视作是属性激活的不同模式;(3) 基于与相关类的直接相似性,有效使用大多数相似类的分类器.但是由于技术的快速发展,这样的分类方式是远不够全面的;并且从本质上讲,这也是因为对零样本学习中“知识”本身的定义不甚清晰造成的.因此,本文基于第 1 节中所定义的知识,在第 2 节中对现有的零样本工作(主要是关于图像分类任务)进行梳理.相比于文献[3],本文中对知

识的定义也更加全面.本文中除“知识”的表述之外,“语义”也是本文的常用表述.而中文的“语义”一词相较于英文“semantic”,往往涵盖内容更加广泛,并且相关文献并没有对“语义”做出准确定义.因此在本综述中,我们结合国内外文献资料,对其进行了一定程度的狭义化:将“语义”狭义化为仅指向“文本的特征表示(例如标签或者属性,这也对应视觉信息经过高度抽象化之后的特征表示)”,它们的共性在于维度较低,且更为抽象.

除了本文主要涉及的图像分类任务之外,零样本学习的策略也运用在很多其他任务领域,诸如识别任务[17−36]、语义分割[37]、图像检索(image retrieval)[38−40]、视频理解(video understanding)[41−44]、检测任务[45−50]等其他视觉任务领域;还有自然语言处理(natural language processing,简称 NLP)[51,52]等文本领域的任务.但是 NLP 领域内的工作相对较少,并且我们可以简单理解为:在现有视觉模型的框架下,对视觉数据的处理转向对文本数据的处理.总体来讲,零样本学习正在渗透到实际应用的方方面面.

本文第 1 节详细阐述在零样本视觉识别任务中,知识的层次以及各自的表示形式.第 2 节介绍基于不同层次知识构建的零样本学习模型.第 3 节则重点阐述目前本领域中一直以来存在的两个主要问题:domain shift 问题和 hubness 问题,并基于问题对现有工作进行总结归纳.第 4 节介绍本领域的通用数据集、评估标准(分类任务)和实验.最后对未来研究趋势进行一定的展望.

1 零样本视觉识别任务中的知识及表示

零样本学习的关键在于“知识”从可见类到不可见类的有效迁移.传统来讲,人们习惯于从知识的存在形式上对知识进行理解和分类.从这个角度来看,知识总体上可以分为 4 种:文本形式、视觉形式、数据分布形式和符号形式.这样划分有助于人们更好地理解不同存在形式的知识本身具备的优缺点,见表 1.

Table 1 Existence form of different knowledge and its advantages and disadvantages

表 1 知识的存在形式以及优缺点

知识存在形式	优点	缺点
文本	拥有较为明确的描述性	缺乏足够的判别性
视觉	拥有较好的判别性	缺乏明确的描述性[53]
数据分布	能够更好的反映数据背后存在的规律	存在某些假设过于严格的情况
符号形式	拥有丰富的专家级别的知识	缺乏合理的知识表示方式以及与特定任务网络的结合形式

其中,文本形式的知识,无论是人类对特定数据集进行的定义还是通过外部获取提取,虽然其拥有明确的描述性,但是由于其数量和维度的限制,以及其中可能存在的噪声,这类知识并没有足够的判别力来对不同类别进行区分.而视觉信息是数据不同类别的真实反映,因此其更具有判别性,但它们中可能包含了更多无法描述的信息(non-robust features)[53],因此无法像文本数据(例如属性)一样有较明确的描述性.数据分布形式的知识,反映了数据集中文本特征或视觉特征的内在规律,更为抽象和高级.从本质上来讲,绝大多数的零样本学习模型所寻求的是不同模态数据之间的对应关系,并寻求这种对应关系的泛化.因此,基于这类更高级特征表示来寻求对应关系的建立,能够使得模型更好地泛化到目标域中.但为了方便地利用这类知识,有的研究者做出了过于严格的假设,例如,“源域和目标域的数据流形结构是一致的”等.而符号形式的知识,知识图谱是其代表,其中包含了丰富的先验信息,例如不同概念层次以及概念之间的显式关系等,但是其难点在于知识采用何种技术手段才能尽可能多地保留其先验内容,即如何选取知识的合理表示形式,还有与深度网络(连接主义)的结合等问题.

传统的知识划分形式无助于人们理解零样本学习领域“知识迁移”的发展历程,因此,本文结合近几年工作,将零样本问题所用数据中蕴含的“知识”定义范畴进行了重新梳理.并根据知识的来源方式,将其划分为 3 个层次:初级知识、抽象知识以及外部知识.采用这样的分类方式,有助于人们更好地理解研究者们如何通过知识的挖掘来缓解甚至克服领域中存在的问题.接下来我们将进行详细的介绍.

1.1 初级知识

初级知识,是指从原有数据集中直接可获取的数据知识,通常包含了类属性知识、类标签知识,以及每条数据的视觉特征.

- 属性(attribute)

属性(attribute)是零样本学习中最广泛使用的知识之一,获得了广泛的关注.在文献[41]中,对属性的定义是指描述某个体或某类所拥有的一系列特性.属性进一步可以分为视觉属性、局部属性.以家猪为对象进行举例,其视觉属性可以是“肉粉色”“皮毛”和“条纹”;其局部属性可以是“4 条腿”“蹄子”和“尾巴”.因此,家猪相对应的视觉属性以及局部属性的表示分别为[1,1,0],[1,1,1].从上面的例子中可以看出:属性向量中的数字均代表了该个体或者类的某一项特性的有无,且数值均为二值(0/1).这类属性知识得到研究者最广泛的使用,被称为人工定义的属性(user-defined attributes),0/1 的数值分布也被称为二值属性向量.

但是,二值属性有一个明显的缺陷,0/1 值并不能表示类间对同一特性的不同强度.例如,家猪和马都有 4 条腿,但是这两个物种二值属性的对应表示均为 1,但视觉上却差异巨大.因此,Parikh 等人[54]提出了相对属性(relative attributes)的概念,即向量的值不仅仅判断特性的有无,并且对应分值的大小还表示该特性的强弱,从而区分不同类别在同一特性上的差异.在一些数据集[17,35]中,分别使用了二值属性和相对属性对同一个目标对象进行描述.

除了二值属性和相对属性之外,还有属性的自动学习[55],即从数据中自动挖掘出相应的属性知识,这排除了人工定义的属性的局限性和不确定性,也被称为数据驱动的属性(data-driven attributes)[3].

最后还有视频属性(video attributes)[23,26,29,48−50],即提取视频中的概念(concept)作为视频属性.

总体来看,属性作为底层视觉像素特征和高层语义特征(即代表了用户对图像的理解,例如类别标签)之间的中间描述层,是对底层特征进行一定抽象的结果,保留了较多信息的同时,也不至于像高层语义特征一样丢失很多描述性信息.因此,属性知识表现出较好的描述性,并进一步具备了较好的共享性和可操作性.例如在 AWA2 数据集中,实现了 85 个属性对 50 个动物类别的描述.但是属性知识需要专家级别的标注,相比简单的类别标注更加复杂和昂贵,这在一定程度上违背了零样本学习的初衷.虽然有属性的自动学习,但是也没有得到很广泛的应用.近两年的趋势表明:更多研究者倾向于其他更加本质的数据知识,例如数据分布等,而非属性知识本身表示手段的革新.

- 类标签(label)

类标签(label)也是零样本学习中最广泛使用的知识之一.以 Mikolov 等人[56]提出词向量的概念为起始点,并伴随着词嵌入技术的普遍使用,这类知识日益受到了研究者们的关注.常用词嵌入模型有 Word2Vec[57]和 Glove[58].获取类标签的词嵌入表示,大致需要进行如下的步骤:首先,使用词嵌入模型在语料中进行训练,例如 Wikipedia Text;然后得到词向量矩阵;最后,通过查询找到类标签对应的词嵌入表示.

通过以上方式得到的类标签知识的优势是巨大的:首先,词嵌入的表示是低维稠密的;其次,词嵌入表示有着很好的空间分布特性,即能够在空间上很好地显示出不同词(每个词代表一个类别)之间的相似性程度.例如,在词嵌入空间中,“狗”和“猫”之间的距离要远小于“狗”和“高楼”之间的距离;还能进一步地作类比和推理,比如“*Vec*(“国王”)−*Vec*(“男人”)+*Vec*(“女人”)≈*Vec*(“王后)”.通过衡量类间的相似性程度,并将其作为一种零样本问题的先验知识,也有助于零样本问题的解决.但区别于前面的属性知识,词向量表示的每一个维度并没有明确含义,这意味着词嵌入之后的类标签表示没有类似属性知识一样明确的描述性,也就失去了知识的共享性这一性质.但在很多研究工作中的模型已经实现了对属性和类标签知识的兼容,即两者的输入在模型中可以相互替换.需要指出的是,由于词嵌入表示是通过词嵌入模型在语料上的训练来获取,而非通过数据集中的多个数据,因此将类标签归入到“初级知识”的范畴中.

- 语义空间中的类原型表示

在属性知识和标签知识中,涉及最多的就是关于“类原型(class prototype)”的表述,是指某个类的代表,而这一代表在语义空间中通常是唯一的,可视作前面属性知识和类标签知识概念上的延伸.在语义空间中,Fu 等人[59]将这一空间中的类原型表示对应为属性向量或者基于类名的词表示,因为在语义空间中,属性表示或者是类名称具备唯一性.因此,很多研究者直接将类标签的词嵌入[11,14,20,28,60−62]或者类属性[13,27,53,59,62−67]作为对应类的类原型,甚至是类别所对应的文本描述[68].其实针对这三者而言,根据数据获取的情况或者是任务的要求,在

绝大多数情况下可以互换[69].除此之外,也有研究者基于这些知识学出了每个类的类原型表示[29,30,41,42,70,71].

- 图像特征(image feature)

图像特征(image feature)也是零样本学习中最广泛使用的知识之一.广义来讲,图像特征表示既可以是指人工设计的特征,例如 SIFT,HOG 等,也可以是图像底层的像素级别的特征,还可以是通过各种深度网络提取的高层抽象的图像特征.近几年,随着端到端深度神经网络框架(例如 AlexNet[72],VGG[73]等)的成熟,当前零样本学习范畴内的图像特征通常是指最后一种.

1.2 抽象知识

初级知识在构建零样本模型时存在较大问题,例如,基于语义空间的类原型表示,由于其唯一性,并不能很好地代表类内方差较大的数据类别;基于初级知识构建的映射,也很容易带来 Hubness 的问题等.在这种情况下,进一步挖掘数据集合所包含的更加高级的知识来减轻或者消除前面存在的问题,就显得很有必要.在这种情况下,研究者通过多条数据甚至整个数据集的总结归纳,挖掘出数据集中的隐藏信息.相较于初级知识,这种类型的知识更能反映出数据的本质特征,我们称为抽象知识.这一范畴通常包含基于图像特征的类原型表示、数据的流形分布以及数据的概率分布.

- 基于图像特征的类原型表示

基于图像特征的类原型表示是前面图像特征的概念延伸,但是其不同于之前基于语义空间的类原型表示,它的表示可以是不唯一的.研究者发现:经过深层网络抽取的图像高级特征,经过 t-SNE 算法[74]的降维并进行可视化,其每个类的数据在空间中呈现簇状分布.因此,有研究者直接将最终提取到的某类所有图像特征[75−78]或者部分图像特征[79]的均值作为该类的视觉类原型表示(前者可以理解为类的质心).但是也有研究者指出[80],基于质心的类原型表示并不能让类间保持很好的判别性.因此,也有模型是通过学习得出的类原型表示[53,80−90].例如,Liu 等人[90]首先基于语义信息来学出类原型表示,然后利用基于均值的图像类原型表示来进行紧致化的结果修正.但是就总体而言,单个类原型不能很好地表示类内方差较大的类别,因此也有工作[83]将基于图像特征的单个类原型表示扩展为多个,其大致过程是将该类包含的所有图像特征,先做聚类,然后将聚类结果的每个簇的均值作为该类的类原型之一,以更好地表示类内方差较大的类别.可以看出,基于图像特征的类原型表示能力远大于基于语义空间的类原型表示.

- 数据的流形分布(manifold distribution)

数据的流形分布(manifold distribution)是数据集中所有(部分)类的整体分布结构,由于视觉特征和语义特征之间存在模态鸿沟,因此,基于不同表示空间获取的数据流形结构可能也是不一致的.为了更方便进行知识的迁移,研究者常常需要一定的假设来对这种情况进行约束,最常见的是假设两个空间中的数据流形分布一致;通过对齐不同空间中的数据结构,或者保持映射过程中数据的结构性,能够有效缓解 Hubness 问题并获得泛化能力更好的模型.有的研究者通过两两类原型之间的欧式距离(Euclidean distance)或者余弦距离(cosine distance)来构建数据的结构图[28,29,53,59,76,88,91−95],这虽然能够直接反映出两个类别之间的相似程度,但是这样简单直接的计算只是基于二维空间分布的前提条件下,一定程度上忽略了数据的分布可能存在更为丰富的流形结构(欧式空间只是流形空间的一个特殊情况).例如,某些对象类会组成超类,并位于相同的子流形上,如果此时再用欧式距离或者余弦距离进行度量,就有可能出现如图 1[14]所示的情况(如果使用欧式距离进行度量,那么 x 将被划分到 z_1 类;如果考虑数据的流形分布结构,那么 x 将被分类到 z_2 类),进一步影响测试集数据的正确分类.因此,有的研究者[13,14,60,87,96]基于更加复杂的数据流形分布,进一步考虑了类原型之间的流形距离(semantic manifold distance).

- 数据的概率分布(probability distribution)

相对于数据的流形分布是挖掘数据集中多个数据的分布规律,数据的概率分布指的是单个数据的生成规律,即通过现有的数据特征,学出这一类特征存在的规律.有的研究者通过建立数据概率分布之间的映射,来使得模型更加鲁棒[30,97−100].伴随着近两年生成式模型的广泛使用,更多的研究者通过生成式的模型来挖掘数据的概率分布知识,并将问题转变为标准的监督学习问题.在基于生成式方法的零样本模型中,通常基于变分自编码

器(variational auto-encoder,简称 VAE)[39,86,101−104]、生成对抗网络(generative adversarial network,简称 GAN)[32,40,44,62,83,99,105−110],或是将两者结合[111−113].其中,VAE[114]优化的是似然下界而非似然本身,而 GAN[115−117]则通过神经网络强大的拟合能力来直接缩小伪数据与真实数据之间的分布差异(Jensen-Shannon divergence,简称 JS divergence).由于 GAN 中对抗训练的存在,使得其训练稳定性相比 VAE 较差.也正是由于对抗训练的思想,GAN 最终生成的伪样本效果整体上要优于 VAE 的表现.但是这类工作均包含了一个隐藏的前提假设:数据服从多元高斯分布.

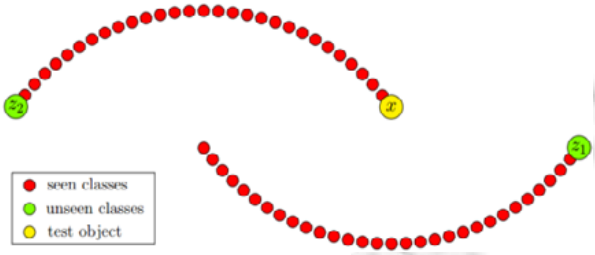


Fig.1 Manifold distance of data in complex manifold distribution[14]

图 1 数据在复杂流形分布中的流形距离[14]

1.3 外部知识

外部知识,顾名思义,是指独立于目标数据集之外的、来自于相关领域的知识,可以被认为是来自于人类的先验知识.传统的深度神经网络得益于大规模标注数据,能够习得有效的层次化特征表示,从而在相关任务领域,尤其是图像领域取得优异的效果.但是也受困于大数据,伴随着数据标注成本日益高昂,单纯依赖大数据的模型性能也已经触摸到天花板,体现出其局限性,比如模型训练过分依赖大数据、模型无法有效与人类先验知识相结合、模型学习结果往往与人的认知规律相冲突(缺乏解释性)等.因此,将外部知识加入到目前零样本问题的解决过程中,可以进一步提升模型在目标任务上的性能(例如模型鲁棒性或者任务相关指标精度).引入外部知识通常有两种形式:其一是引入与类别相关的、除属性和类标签之外的其他来源的描述,这一过程类似于“数据增广”,即在现有数据集基础上,进一步扩大数据来源;其二是直接利用现有大型知识库,主要存在形式为知识图谱,是人工智能符号主义的典型代表.

- 类别文本描述(text-description)

在数据集中,由于图像对应的文本信息,例如类标签、属性等知识包含的信息是有限的,这在某种程度上限制了模型在 ZSL 任务上进一步提升性能.因此,有研究者通过扩展数据集中文本信息的来源,即增加数据量来增加所包含的知识量,从而进一步提升模型处理对应任务的性能.额外数据来源有很多,例如,可以用从网站(例如 Wikipedia 或 Wikipedia articles)的词条中或者对应的专业领域网站,获取到针对该类更多的描述[107,118−120].同样,也可以通过搜索引擎[47]等其他渠道.在挖掘到额外的文本描述之后,通过一些自然语言处理(natural language processing,简称 NLP)技术,例如传统的词袋模型(bag of words,简称 BOW)[119]或者提取 TF-IDF 特征[118,120],对这些信息低维嵌入进行处理;还可以利用词嵌入+深度模型等方式,对额外的信息进行编码,映射到一个低维的表示空间中[107].需要指出的是,在获取额外知识的同时,也需要过滤掉其中包含的噪声.如何有效过滤噪声并同时保留任务相关知识,是目前比较棘手的问题.

- 知识图谱(knowledge graph)

知识图谱的本质是语义网络,是一种图结构的数据,由“节点-边-节点”组成.其中,节点代表“概念”或“实体”;边则代表两个节点之间的关系,用来描述现实世界中的概念、实体记忆以及他们之间丰富的关联关系(知识谱图发展报告 2018).在零样本学习领域,常用的知识图谱有 WordNet,ConceptNet[121,122]等.想要利用知识图谱,首先要解决的问题就是如何对知识图谱进行合理的表示.由于知识图谱中的实体、概念以及关系均采用了离散的、显式的符号化表示,而这种表示形式难以直接应用于基于连续数值表示的神经网络中,因此,将其包含的知识尽

可能地嵌入表示在一个低维向量空间中,是知识图谱与深度神经网络相结合的前提条件.在这方面,有两类主要方法:以翻译模型为代表的传统知识表示技术[123];以图神经网络(图神经网络(graph neural network,简称GNN)[124]和图卷积网络(graph convolutional network,简称GCN)[125]是代表性的两种)为代表的深度知识表示技术.尤其后者的出现,使得知识图谱的表示学习跨入了深度学习的领域,将以知识图谱为代表的符号主义和以深度学习为代表的连接主义走上协同并进的道路.其次,还有另一个关键的问题在于"符号主义和连接主义,两者分别包含的人类先验知识和从数据中学出的经验知识,怎样结合才能有效提升特定任务性能?".以 WordNet 知识图谱为例,主要包含了两方面可用的知识:其一是层次化的概念表示,例如"哺乳动物-猫科动物-东北虎",有研究者利用这一层次化的知识表示形式来引导判别性特征的生成[40,126−128];其二是包含不同类之间的显式关系,有的研究者利用这一特性来辅助源域到目标域之间知识的迁移[31,46,62,129−132].需要指出的是,后者还处在起步的阶段,也是如今的"数据+知识驱动"思想潮流的主要呈现形式.

1.4 知识之间的联系

初级、抽象、外部这 3 个层次的知识并不是孤立存在的,它们之间也存在千丝万缕的联系.基于第 1.1 节~第 1.3 节的介绍,它们之间的详细关系如图 2 所示(抽象知识中的数据分布包含了数据的流形分布和概率分布).

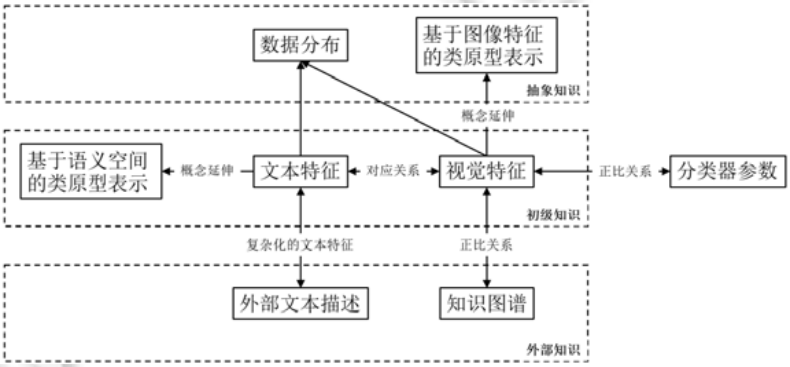


Fig.2 Knowledge relation framework

图 2 知识关系架构

从图 2 中可以看出:初级知识中的文本特征与视觉特征存在对应关系,即某个类的标签或者类属性对应该类全体图像样本;而基于语义空间的类原型表示则可以看成文本特征在概念层次的进一步延伸.基于初级知识中的文本和视觉特征,我们可以建模它们各自的多个数据点之间的流形分布,或者学出他们的特征概率分布;而基于图像特征的类原型表示也可以看成是视觉特征在概念层面的进一步延伸,多类原型表示也可以在一定程度上反映类内方差的信息.外部知识中的外部文本描述,则可视作复杂化的文本特征,因为它在包含更多信息的同时,也包含了更多的噪声,使得任务相关信息的处理和提取变得更加困难.

除此之外,我们在原先不同类型知识之间关系的基础上,新增加了两个关系.

(1) Zhou 等人[76]提出了一个假设,即"如果两个类视觉上相似,那么其分类器的参数也相似".因此,我们可以通过这一假设,将分类器参数加入到语义空间和视觉空间的对应关系构建过程中,代表性工作如文献[76,130,131]中的 ZSL 模型.

(2) 针对视觉特征和知识图谱之间的关系,Deselaers 等人[133]基于对 ImageNet1k 数据集的分析得出一个结论:类中图像视觉相似性与类标签的语义相似性在总体上成正比(总体上).即在人类定义范畴内的概念之间的相似性,能够反映在这些概念内所包含的图像之间的视觉相似性上.该文献类标签之间的相似性度量使用的是 JC 距离(Jiang and Conrath semantic distance,简称 JC distance),因此更确切地说,视觉之间的相似性和所对应的类标签在知识图谱中的显式距离,在总体上成正比.

2 基于知识迁移的零样本视觉识别模型

在这部分内容中,我们将基于本文第 1 节所定义的不同层次的知识,对现有零样本学习的相关工作进行梳理.为便于读者理解该层次模型使用知识的方式,本文在每一层次的模型中,进一步进行了相应的合理划分.需要指出的是,这一过程是向下兼容的,即依据模型所用到的最高级层次知识进行模型的划分.例如,模型如果同时使用了初级和抽象层次的知识,就将其归纳到基于抽象知识的模型范畴中.接下来,我们将进行详细的介绍,介绍的重点将是最具代表性的图像分类任务领域的模型.

2.1 基于初级知识的零样本模型

在零样本学习领域,大部分的工作仅仅使用了数据集中包含的一些初级知识(属性、类别标签、视觉特征等).在这类模型中,属性作为标签和视觉特征之间的中间描述,在源域类别和目标域类别之间拥有良好的描述性和迁移性.因此,有很多研究者从概率的角度去进行属性学习的工作.其中,Lampert 等人[33,63]提出了代表性的属性学习模型,掀起了属性学习的热潮,后面很多工作均是受此启发.除了属性学习之外,也有的研究者将文本特征空间和视觉特征空间之间进行映射建模,这样更接近零样本学习的本质.深度学习的兴起,其强大的拟合能力极大地提升了模型性能,这也让很多研究者使用深度模型去重复这些建模思想.下面,我们分别对基于属性迁移的模型和基于映射的模型两大类方法进行介绍.

2.1.1 基于属性迁移的模型

如前面第 1.1 节中所述,属性知识作为一种中间描述,能够让可见类和不可见类之间实现信息的共享,具备了良好的迁移性.因此,很多研究者将属性作为底层特征和高层抽象特征(即标签)之间的中间表达层,进行零样本的学习.根据建模的方法,大致可以分为概率模型和深度模型.

- 概率模型(属性学习)

Lampert 等人[33,63]首先提出了两个具有影响力的基于属性的概率模型:直接属性预测模型(direct attribute prediction model,简称 DAP)和间接属性预测模型(indirect attribute prediction model,简称 IAP).

DAP 模型的架构如图 3 所示.

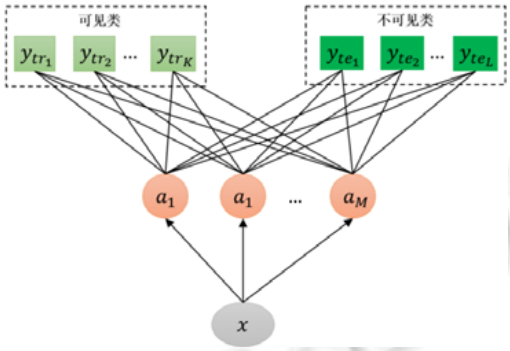


Fig.3 DAP model

图 3 DAP 模型

图 3 中,x 表示为可见类样本的图像底层特征,$a_1,a_2,\ldots,a_M$ 表示为可见类与不可见类之间共享的 M 个属性,顶层的 $y_{tr_1},y_{tr_2},\ldots,y_{tr_K}$ 和 $y_{te_1},y_{te_2},\ldots,y_{te_L}$ 分别代表可见类的标签和不可见类的标签,K 和 L 分别表示可见类和不可见类的类别数.在训练过程中,首先将可见类样本的图像底层特征和二值属性向量作为输入,然后利用支持向量机训练属性分类器,最终得到样本和属性之间的关系 $p(a_M|x)$.在测试过程中,结合之前得到的属性分类器,首先输入不可见类样本的图像底层特征,即可实现对该样本的属性预测;然后利用类-属性矩阵(即类别和属性之间的先验知识),最终完成对测试样本的分类;测试时,对测试样本的属性进行预测,再从属性向量空间里面找到和测试样本最接近的类别.从公式角度来看,在测试时,首先计算不可见类样本属于每个未知类的概率:

$$p(y_{te}\mid x)=\sum_{a\in\{0,1\}^M}p(y_{te}\mid a)\cdot p(a\mid x)=\frac{p(y_{te})}{p(a^{y_{te}})}\prod_{m=1}^{M}p(a_m^{y_{te}}\mid x) \tag{1}$$

其中,$p(y_{te})$为每个未知类的先验概率.因此,利用了最大后验估计方法(maximum a posterior,简称 MAP),其预测最终不可见类标签的公式如下:

$$f(x)=\arg\max_{k=1,\ldots,L}p(y_{te}\mid x)=\arg\max_{k=1,\ldots,L}\prod_{m=1}^{M}p(a_m^{y_{te_l}}\mid x) \tag{2}$$

有研究者基于 DAP 模型的思想框架进行了进一步深入的工作.针对属性本身,区别于文献[33,63,64]均是专家标注的属性,有的研究者进一步地扩展了属性的来源[20,41,42,134,135],还有的研究者[136,137]在建模时考虑了不同属性的重要性程度.Jayaraman 等人[138]则注意到了属性在预测不可见类时的不可靠性(unreliable)问题,并用随机森林的算法对其进行了处理,通过统计每一个属性在预测时的错误率,来提高属性预测的鲁棒性.Rohrbach 等人[139]则借助了知识库来降低人工标注属性的成本.针对模型本身,Huang 等人[64]将属性学习转化为了超图分割的问题.在超图中,每个节点表示一个样本,每条超边表示样本共享的属性.文献[41,42]使用了主题模型来替代 SVM.Yu 等人[140]使用了作者-主题模型来建模特征-属性分布.Wang 等人[141]则使用了统一的概率模型去建模目标独立属性和目标依赖属性之间的关系.此外,在其他类型任务中,Hariharan 等人[142]将 DAP 模型进一步扩展到了多标签分类领域.Cheng 等人[21]将这一原理扩展到了动作识别领域,具体是将动作转化为属性特征然后加入到零样本网络训练当中.

IAP 模型的架构如图 4 所示.

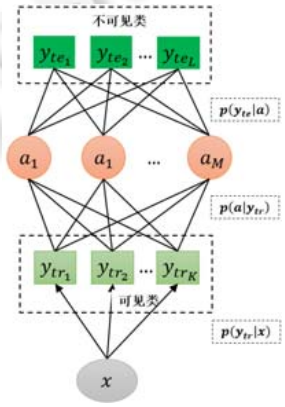


Fig.4 IAP model

图 4 IAP 模型

图中参数的定义与图 3 一致,最大的不同之处在于:IAP 模型使用已知类的标签,间接地学习图像底层特征到属性的映射.在训练过程中,首先将可见类样本的底层特征和其对应标签作为输入;然后利用支持向量机训练出类别-特征模型,得到样本和每个可见类之间的关系$p(y_{tr}|x)$;接着,根据类别-属性关系矩阵获得属性-类别模型;最后,将前两个模型进行结合,推导出属性预测模型.在测试过程中,输入不可见类样本的图像底层特征,首先实现对该样本的属性预测,然后利用类别-属性矩阵并结合贝叶斯定理,实现对该样本的标签预测.从公式的角度来看,在测试时,首先计算不可见类样本属于不同属性的概率:

$$p(a_m\mid x)=\sum_{k=1}^{K}p(a_m\mid y_{tr_k})\cdot p(y_{tr_k}\mid x) \tag{3}$$

其中,$p(a_m\mid y_{tr_k})$为属性-类别模型.然后结合属性-类别模型,进一步预测样本的类标签.这一过程同样利用了 MAP 方法,其预测不可见类标签的公式如下:

$$f(x)=\arg\max_{k=1,\ldots,L}p(y_{te}\mid x)=\arg\max_{k=1,\ldots,L}\prod_{m=1}^{M}p(a_m^{y_{te_l}}\mid x) \tag{4}$$

相较于 DAP 模型,IAP 模型在实际中使用较少,但也有学者基于 IAP 模型进行了深入的研究.Kankuekul 等人[143]考虑到 IAP 模型相较于 DAP 模型拥有更低的计算成本等优势,因此基于 IAP 模型提出了一种在线增量学习算法,能够动态地学习新的属性以及更新现有的属性.

- 深度模型

在基于属性迁移的模型(属性学习)中,Morgado 等人[144]使用了深度神经网络,将 RIS(recognition using independent semantics)方法和 RULE(recognition using semantic embeddings)方法(该文献中的分类定义)相结合,即利用两者的互补性,将 Deep-RIS(属性分类器)加入到 Deep-RULE 的过程中,取得更好的效果.Lu 等人[145]则为每个属性训练一个单独的深度属性分类器,并进一步筛选出置信度高的属性组合成为不可见类的伪样本进行训练,将 ZSL 问题转化为了一个有监督学习的问题.

- 针对属性特性的学习

除了上述的方法之外,有研究者从属性本身的角度出发,就“属性应该具备怎样的特性”这一问题进行了深入探究.Wang 等人[141]提出了一个统一的贝叶斯概率模型,发现和捕获 Object-Dependent(例如“条纹”和“白色”依赖于斑马,互斥于北极熊)和 Object Independent(“翅膀”和“爪子”与许多不同的鸟类有关)的属性关系,这是 DAP 和 IAP 类的模型所没有考虑的问题.Jayaraman 等人[146]针对之前模型中存在“属性的应用依赖于正确相关语义概念”的问题,提出了一种多任务学习方法(每个属性被视作一个任务).该方法通过将相似的图像特征与含义相近的属性对齐,不相似的图像特征与含义较远的属性对齐,并强制结构稀疏,最后为每个属性学得一个分类器,这个分类器能够更好地处理经常协同出现的属性,而非将它们合并为一个属性.Liang 等人[147]注意到了“同一属性在不同的视觉空间中变化很大”这一问题,将类别标签信息和图像特征共同映射到一个共享潜在表示空间,然后进行进一步的属性分类器学习,使得最终学出的分类器依赖于特定而非所有类共享的类信息.Gan 等人[148]聚焦于属性学习中一个最基础的问题:如何抽取出更具泛化性的属性?作者认为属性检测是一个多源域(该文献中每个类被当作是一个域)泛化的问题,并利用现有的分类器最终获取了高质量的属性检测器.Jiang 等人[71]基于字典学习的方式,将人工定义的属性进行重新组合,学习了潜在属性表示.但从另一个角度来看,潜在属性的描述性没有人工定义的属性明确.

除此之外,Li 等人[149]将 DAP 和 IAP 方法进行了不同方式的结合.Zhu 等人[84]认为,在映射过程中应该保留视觉中包含对应语义空间的部分,因此将视觉样本嵌入到一个低维的概率矩阵中(具体作用是衡量对象中某个属性的出现概率(语义组件)),在视觉空间、属性的语义空间和标签的语义空间之间构建了更加精确的关系.在其他类型任务中,Kumar 等人[17]将属性用于人脸的识别,具体提出了两种类型的分类器:前者筛选出 65 个可描述性的属性用于分类器的训练;而后者则不依靠人工标注,转而依靠图像区域的相似性来进行人脸的识别.

2.1.2 基于映射的模型

基于属性迁移的模型,其大致构建了“图像特征-属性-类标签”这 3 个层次的架构.而基于映射的模型从不同模态数据对齐的角度出发,直接构建了“视觉特征-文本特征”两层次的架构,即构建视觉空间和语义空间之间的映射,并且要求映射是具有泛化性的(从源域可见类到目标域的不可见类).这种映射形式从方向到形式上都是多种多样的,我们根据所构建映射的方向差异,将这类模型分成以下 4 类:从视觉空间到语义空间的映射(正向映射)、从语义空间到视觉空间的映射(反向映射)、视觉空间和语义空间的双向映射、视觉空间和语义空间中的数据共同映射到共享潜在空间(共同映射).下面分别对它们进行介绍.

- 正向映射

根据映射方式的不同,这种映射方式既可以是线性的,也可以是非线性的.其基本形式如下所示:

$$\min\|f(X_{tr})-Y_{tr}\| \tag{5}$$

其中,$f(\cdot)$是指正向映射函数,最基本的,它可以只由一个映射参数 W 组成;Y_{tr} 可以是标签,也可以是属性.研究者还进一步地加入了各种类型的正则化项,例如常见的二范式约束.针对映射函数 $f(\cdot)$,有的研究者[61,82,150,151]致力于扩充映射内容丰富性,例如 Yu 等人[150]增加了类分离性损失、Li 等人[82]增加了潜在属性空间学习等;而针对映射函数本身,有的研究者[19,23,81]利用神经网络强大的拟合能力来进行映射、有的研究者[65,152−154]通过使用矩阵

分解的技术来构建更加细粒度的映射.

损失函数 min 的方式也分多种,例如有均方误差的形式,也有基于均方误差的 Triplet Loss[45,65,78,155]的形式.除了岭回归形式的损失之外,基于概率的损失(例如深度神经网络模型中常见的交叉熵损失)也可以被认为是存在于正向映射过程中的损失方式.Atzmon 等人[156]基于 OOD(out-of-distribution)的思想构建了一个概率模型框架，用以区分来自于可见类和不可见类数据,并通过 softgate 的方式将可见类分类器(expert model for seen classes)和不可见类分类器(ZSL expert)结合起来,以更好地适用于广义零样本任务(GZSL),该框架可以整合任何输出为类别概率的模型.Zhu 等人[78]通过利用语义信息引导的多注意力机制来定位图像中最具判别性的部分,并在 Softmax Loss 和 Triplet Loss 的共同作用下,挖掘出类间分离、类内紧致的视觉特征.

在测试阶段,由视觉到语义空间的映射完成后,使用 k 近邻分类器来对不可见类样本进行认知.在其他类型任务中,文献[19,23,45,157]则分别将正向映射的思想扩展到了零样本面部欺骗攻击任务(face anti-spoofing)、零样本的动作识别、目标检测等任务当中.其中,Liu 等人[19]引入树状的条件 CNN 结构来进行零样本面部欺骗攻击任务(face anti-spoofing);Jain 等人[23]利用神经网络将视觉内容映射为目标概率,并利用词嵌入技术来建立在目标和动作之间的联系;BANSAL 等人[45]将从图像目标中提取到的图像特征,通过相似性比较的策略进行类别的认知.

- 反向映射

一般来讲,视觉空间维度要比语义空间维度大,所以建立从视觉空间到语义空间的映射往往会丢失判别性信息,产生特征空间坍缩.因此相较于正向映射,反向映射能够保留更多的描述性信息,从而能够防止特征空间坍缩,进而缓解零样本学习中存在的 Hubness 的问题.其基本形式如下所示:

$$\min\|X_{tr}-g(Y_{tr})\| \tag{6}$$

其中,$g(\cdot)$是指反向映射函数,其设定类似于正向映射的函数;Y_{tr} 可以是标签,也可以是属性.研究者们还进一步地加入了各种类型的正则化项,例如常见的二范式约束.min 的方式也是多种多样的,例如均方误差的形式.在这类方法中,目前只有少量的研究[11,68,89,158,159],更多的是针对映射函数的改进,例如,Kodirov 等人[11]增加了源域和目标域所学字典的相似性约束,使得学得的映射更加具有泛化性;Changpinyo 等人[89]直接建立视觉类原型(簇中心)与语义类原型之间的反向映射.

在测试阶段,由语义空间到视觉空间的映射完成后,使用 k 近邻分类器来对不可见类样本进行认知.

- 双向映射

这种映射方式同样也是为了解决正向映射过程中产生的信息丢失、特征空间坍缩的问题.为了保留更多的判别性信息,研究者将映射到语义空间中的特征,再重构回视觉空间.这样,学习到的映射就能够得到保留更多的信息.其基本形式如下所示:

$$\min\|X_{tr}-g(Y_{tr})\|+\|f(X_{tr})-Y_{tr}\| \tag{7}$$

其中,$f(\cdot)$和 $g(\cdot)$是指映射函数,这两个映射函数同样可以是线性的,例如只是互为转置的映射参数 W,也可以是非线性的;而 Y_{tr} 可以是标签,也可以是属性.min 的方式一般为均方误差的形式.

Kodirov[66]构建了基本的自编码器(auto encoder,简称 AE)结构来实现双向的映射,其中,编码器(encoder)将视觉空间映射到语义空间(属性),然后再重构回去(decoder).基于这样的思想,许多后续的研究针对双向映射函数进行了改进.Annadani 等人[160]在加映射层数的基础上加入了类别之间的关系约束,且映射方向与文献[66]相反.Wang 等人[101]使用了生成式的模型 VAE 来构建双向映射,将语义空间作为隐层(将属性假设为高斯分布,即属性经过两个线性映射分为生成均值和方差).这样能够更好地揭示数据的复杂结构,从而学出判别性更强的特征表示.Zhao 等人[12]利用双向映射来生成域不变的特征.此外,有研究者扩充了映射方式的丰富性.Lu 等人[67]构建了竞争性的双向映射(competitive bidirectional projection),在构建双向映射的基础上,先利用不可见类与可见类之间的相似性关系来辅助不可见类伪样本的生成,再通过 Competitive Learning 机制,使得伪样本离最相似不可见类中心(只是将视觉特征空间中,映射过来的语义特征作为类原型)最近,离次优的中心较远,使得生成的模型更加泛化和鲁棒.Chen 等人[161]在实现上述过程(重构)的同时,将语义空间分解成两个子空间(两个子空间分

别进行分类和重构任务,可以认为是两个互相冲突的任务).通过对这两个子空间进行对抗学习,使得学出的嵌入表示既能保留细节又具有判别性.Bin 等人[162]在构建双向映射的同时,将所提取的特征分解为语义(描述性)特征、非语义(非描述性)特征以及非判别性特征.通过这种方式来提取出更加具有判别性和泛化性的特征,从而增加模型的泛化能力.

这类方法在测试阶段对不可见类的认知类似于正向映射.

- 共同映射

基于共同映射的模型,其形式更加多样,我们下面分别进行介绍.其基本形式如下所示:

$$sim(f(X_{tr}),g(Y_{tr})) \tag{8}$$

其中,$f(\cdot)$和 $g(\cdot)$分别指语义空间和视觉空间到共享空间之间的映射,这一映射函数既可以是线性的,也可以是非线性的;Y_{tr} 可以是标签,也可以是属性;而 $g(Y_{tr})$既可以是映射,也可以是指类名对应的词嵌入的表示.

共同映射形式的多样性体现在 $sim(\cdot)$的多样性上,$sim(\cdot)$主要的存在形式是兼容函数(compatibility function),两个向量直接内积相乘($f(X_{tr})\cdot g(Y_{tr})$)的形式是主要模式之一[10,52,159,163−168].其中,Frome 等人[159]提出了著名的 DeVise 模型,将 CNN 提取的图像特征和标签的词嵌入表示进行内积形式的相似度计算,然后使用度量学习的 Ranking Loss(triplet loss)将它们学习到一个共享潜在表示空间.文献[165,166,168]则是这一模式的基本延续.文献[163,164]则在此基础上考虑了语义的组合问题.文献[10,167]采用了集成学习(ensemble strategy)的思想.其中,文献[10]与字典学习相结合,构建了多个字典,从而能够更好地重构可见类与不可见类所共享的潜在语义字典.文献[167]则通过最大化可见类标签矩阵和随机选取的不可见类标签子矩阵之间的关联性来产生多个标签映射权重,进而映射出多个标签子矩阵;同时,相对应的视觉特征提取模块也产生同等数量的分支,然后两个模块对应分支通过内积相乘(对齐的目的)并通过集成的标签预测方式来为不可见类数据进行打上高置信度的伪标签,迭代地加入到训练过程中.

Yazdani 等人[52]将共同映射的思想扩展到了 Spoken Language Understanding 任务中,直接构建了句子与标签之间的相似性.

双线性兼容函数($f(X_{tr})^{T}Wg(Y_{tr})$)也是另一种兼容函数的主要形式,这一形式的目的是学出来自两个空间表示的最大兼容分数(maximum compatibility score)[27,95,169−175].其中,Yu 等人[173]为了解决在映射过程中各个样本可靠性(贡献)不同的问题以及 Domain Shift 问题,提出了 ASTE(adaptive structural embedding)和 SPASS(self-pased selective strategy)方法,在构建映射的同时,前者自适应地调整松弛变量,以体现训练实例之间的不同可靠性,使得映射更具辨别性;后者通过迭代地迁移可靠性逐渐减弱的不可见类样本以缓解 Domain Shift 问题,同时极大地缩短了训练时间(也用了矩阵分解技术).Jiang 等人[174]在训练过程中采用了自适应的方式,即加入不可见类的文本数据来缓解可见类与不可见类之间存在的 Domain Gap 问题.Song 等人[175]在构建共同映射的过程中,除了将可见类数据进行准确的映射约束之外,还将不可见类的数据强制映射到其他点,从而缓解了 Domain Shift 问题.在其他类型的任务中,Wang 等人[27]将双线性映射方式扩展到了零样本动作识别领域,其模型框架类似于文献[159].

除此之外,$sim(\cdot)$也可以是均方误差[176]或者是余弦相似度等形式,也可以是基于这些基本相似性度量方式进一步构建的 Triplet Loss 形式.Tsai 等人[177]在视觉分支和文本分支映射的过程中加入了 AE 的结构,并在隐层施加了分布对齐的约束,然后分别将 AE 的隐层映射到共享的潜在表示空间.

- 其他基于映射的模型(多种映射方法混合)

这个类型的映射方式,其主要借助判别性损失(例如 softmax loss)来完成正向(或者反向)映射,并在此过程中加入对方模态(除正在进行映射模态之外的另一种数据形式)的信息.我们依据模型的构建思想,可以分为传统模型和元学习模型.

在传统模型中,Liu 等人[69]在映射的过程中,通过温度校正(temperature calibration)[178]来缓解由于在可见类数据上的过拟合导致的对可见类的域偏移现象,最终将两种模态的信息映射在同一空间.Jiang 等人[179]定义了可见类的分类损失与不可见类的迁移损失,将知识迁移的过程进行了一定程度的量化,使得提取的特征同时具

有判别性和迁移性(正向).Liu 等人[70]之前的模型均是基于空间中可见域和不可见域的数据分布在样本级别具有一致性这一假设,而该假设过于严格,因此,Liu 等人提出不寻求在样本级别上进行映射,转而致力于任务级别的一致性,即,以任务为基本单位来构建不同空间之间的映射关系(任务是指对数据集的不同划分).具体来说,对可见类进行不同的划分,形成N个任务;然后,每个任务中类的属性值通过非线性的方式转化为类原型,并与该任务中的图像进行相似性度量(PEC),提升类原型在可见类样本中的泛化性;最后,将相似性度量的结果进行归一化表示之后,完成分类(cross-entropy loss,简称 CEP),其本质上是通过训练和测试在任务层面的对齐,使得训练阶段尽可能地仿真模型的测试环境.模型的思想如图 5 所示(每个几何形状表示一个样本,每种颜色表示一个类).

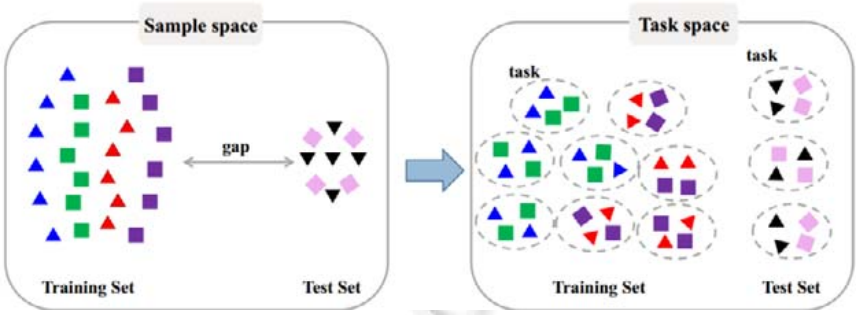


Fig.5 Core idea of CPL (convolutional prototype learning)[70]

图 5 CPL(convolutional prototype learning)核心思想[70]

元学习模型区别于传统的数据集划分方式,对数据集作了进一步的划分,即进一步将训练集划分为两个部分(这一划分的标签集情况,不同文献中,根据不同的训练要求有着不同的要求);此外,元学习采用了 Episode-Based 的训练策略,即其每次迭代都会随机抽取所有类别的子集作为一个训练任务,这样的目的均是为了最大程度的追求模型的泛化性能.Sung 等人[180]提出的基于元学习的零样本学习模型将整体数据集分割为训练集、支持集(support set)和测试集,并且其支持集和测试集共享相同的标签.该模型的核心思想是,通过比较的方式来认知新的事物.因此,将训练集的数据进一步随机划分为样本集(sample set)和查询集(query set)(这样的划分用于仿真测试阶段的 support set/test set).具体做法是:将样本集分别与查询集中的样本语义特征作拼接,然后输入到分类网络中进行训练.而 Hu 等人[181]提出的模型将训练集进行了随机划分,并要求两者的标签集是互斥的.该模型的核心思想是,根据零样本学习中两个域相似性的差异做出不同程度的修正.因此,所提出的模型包含两个模块——Task Module(learner)和 Correction Module(meta-learner):前者将语义特征作为输入并输出初始的预测(视觉特征的质心);后者将前一模块的预测结果、可见类数据、不可见类语义特征作为输入,输出修正量,目的是对前一模块的结果进行修正.与前一模块的输出相加,即为最终的预测结果.

除了图像分类的任务之外,Shen 等人[38]将这一思想用于基于轮廓的图像检索任务,除了各自模态数据的编码网络之外,还利用了图卷积网络和 Kronecker 融合来增强两种模态数据(草图和真实图像)的一致性,并最终用于基于轮廓的图像检索任务(哈希检索).

总体来讲,基于初级知识的模型,所使用到的知识范畴包含了类别标签、视觉特征、属性这 3 种.在这 3 类知识中,类标签虽然包含的信息没有属性丰富,但是其词嵌入的分布式表示特性(该表示形式还具备一定的推理能力,这也是类别标签的词嵌入表示形式所隐含的知识)也能较好地完成零样本学习任务.文献[170]证明了类标签词嵌入的表示形式和属性在表示能力上基本上是等同的.因此,在基于映射的零样本学习模型中,其语义空间中的属性知识和类别标签知识在大部分时候可以进行互换;同时,也有模型可以兼容多种形式的映射[51,182].但是,基于映射的模型无法很好地反映数据类内方差的特性[62],并且基于共同映射的方法有两个局限性:一是不能使用高效的判别分类器,二是不能有效地处理 GZSL 任务[99].

2.2 基于抽象知识的零样本模型

在零样本学习领域,初级知识仅代表可以直接获取的信息.而随着研究的深入,研究者进一步借助初级知识挖掘出了基于图像特征的类原型表示、数据流形分布、概率分布等更加高级的抽象知识,这些知识能够更全面地反映出数据的本质特征,从而能够学出泛化性能更好的模型.因此,接下来我们将从基于视觉类原型的模型、基于数据流形分布的模型和基于数据概率分布的模型去分别展开阐述.

2.2.1 基于视觉类原型的模型

基于视觉类原型的模型,是将所挖掘出的类原型表示作为一个中介,然后去进一步地构建模型,如用于推理等,而非直接用于认知不可见类.从这点来看,其用法更加近似于初级知识.

Zhao 等人[85]使用了基于类内均值的视觉类原型表示来构建映射模型,并与基于合成视觉类原型表示的模型进行了对比.Wang 等人[80]使用基于交叉熵(cross-entropy loss)的损失学出每个视觉类的类原型表示,这样学出的类原型相较于基于质心的类原型表示,能够让类间保持足够的判别性.Changpinyo 等人[87,89]基于视觉类原型提出了样本合成模型,具体来说,在其视觉分支,为了获取每个类的聚类中心表示,为所有样本做 PCA,然后加和求均值(在该文献中被称为视觉特征代表),从而将视觉特征从视觉空间变换到语义嵌入空间(该文献中语义嵌入空间可以理解为共享表示层),最后使用支持向量回归机 SVR(多核回归模型)来建立从语义空间到语义嵌入空间的映射,并最终基于对模型不同的理解.提出了两种方式对不可见类的认知手段:首先,通过训练好的模型得到预测原型(examplars)的表示,如果将预测的原型视作训练数据,那么直接利用训练好的模型,共同映射到语义嵌入空间进行最近邻的分类;如果将预测原型视作改进后的语义特征表示,则可以整合入任何现有的 ZSL 框架中,辅助零样本认知过程.前面的模型均基于“源域和目标域数据分布一致”的假设,而这一假设过于严格.Wan 等人[77]利用目标域样本进行聚类得到每个类的类中心,然后同时缩小不可见类样本与模型训练得到的类中心和聚类得到的类中心的距离,并利用二部图匹配对这一机制进行优化(两个虚拟的图进行一对一对齐,但没有上升到流形对齐的高度).该模型的反向映射机制和直推式的机制均有利于缓解 Domain Shift 现象(该文献事实上使用了基于视觉的类原型表示,但是没有 Prototype 这一表述).

如本文第 1.4 节所述,分类器参数也可以认为是正比于类原型的表示,因此很多模型也是基于分类器进行进一步操作的.Misra 等人[183]利用了分类器组合的思想,认为复杂视觉概念是简单概念进行组合的结果,并且进一步的认为:应该用属性视觉分类器和目标视觉分类器组合得到新的复杂视觉分类器,例如“红色(属性)+酒(目标)=红酒”.具体过程为:通过分类器参数组合,然后输入到转化网络中,计算与真实目标的内积,从而进行训练.

2.2.2 基于数据流形分布的模型

研究者主要利用数据流形来主要达到两个目的.

(1) 由于在基于映射的模型中,跨空间的数据映射为使重构误差最小化,从而倾向于学习两个空间数据之间的共性,这会导致数据的判别性出现不足.鉴于这种情况,有的研究者将流形正则化项加入基于映射的零样本学习过程中,保持数据的结构,从而增加映射的泛化性.

(2) 不同空间进行的映射所产生的 Domain Shift 问题,其本质上可以视作模态鸿沟(media gap)的问题,从数据分布的角度,将不同空间的数据进行流形对齐的操作,实质上就是缓解这一问题(文献[80,94]明确提到这一点).

下面,我们根据研究者是否考虑数据的复杂流形分布,对相关工作分别进行详细的介绍.

在零样本学习中,模型的泛化性能显得尤为重要.而在半监督学习中,增加流形正则化项(manifold regularizer),能够增加对测试数据的泛化能力[184].其基本形式如下所示:

$$\sum_{ij}^{|Y_{tr}|} w_{ij} \| f(X_{tr}^i) - f(X_{tr}^j) \|_2^2 \tag{9}$$

其中,X_{tr}^i 和 X_{tr}^j 是指不同的数据,$f(\cdot)$是指视觉空间和语义空间之间某种形式的映射,w_{ij} 为两个节点的相似性权重.需要指出的是,流形正则化项不仅可以应用到语义空间和视觉空间,还可以进一步地将范围由可见类有标记数据扩展到不可见类无标记数据当中.

- 简单数据流形

仅考虑简单数据流形结构(pair-wise 级别)时,不同的研究者有着不同的正则化项构建方式.Xu 等人[28]在映射模型的基础上,通过使用训练数据和测试数据共同构建 kNN 图来添加流形正则化项约束.文献[88,93]则通过构建潜在空间的图正则化项来保持数据的几何结构:前者还将 DAP 模型的图像特征层和属性层进行了交换,使得模型能够生成一些不可见类的样本;后者则将每个样本被视作可见类的分数组合.Xu 等人[92]用两个流形正则化项来分别保持数据映射过程中在视觉特征空间和属性空间中的几何结构.在其他类型任务中,Qin 等人[29]将这一思想运用到了零样本的动作识别领域,使用了从语料中训练得出的语义表示而非属性特征来判别更加细粒度的动作场景,并在此过程中使用类级别的语义相似性矩阵来保持数据结构,进而维持判别性.

除了增加流形正则化项来保持映射过程中数据的判别性之外,Zhang 等人[91]还提出了结构化的预测方法,即通过最大化后验估计来获取目标域数据的分布,使得潜在空间中对不可见类数据的标签分配是平滑的.Wang 等人[94]简化了文献[91]中复杂的结构预测过程,首先建立从视觉空间到潜在空间的映射,相同一批数据在潜在表示空间的结构需要与其在语义空间中的结构保持一致,从而得到不可见类在潜在表示空间中的节点表示.Jiang 等人[53]提出了一种双字典学习方法,通过视觉空间和语义空间的类原型对齐来使得数据的结构对齐,目的是利用视觉空间的判别性来提升语义空间中的判别性.

有的研究者则直接从特征提取的角度出发,去考虑如何更好地保持特征空间的结构.Li 等人[95]考虑到数据可能存在的类间方差小以及类内方差大的情况,增加了图像特征结构约束来归一化类内和类间样本的距离,使得提取的图像特征能够保持空间结构.这一工作虽然考虑到了所提取特征的空间结构的保持,但是并未上升到数据流形的高度.Wang 等人[80]将排序损失和结构优化损失相结合,在学出的共享表示层中,除了保持不同模态数据的对齐之外,同时也能够确保空间中的视觉特征表示拥有更好的结构,使得判别性更好.

- 复杂数据流形

前面的方法均仅考虑了数据的简单流形结构,事实上,数据中存在着更为复杂的流形结构,不同的研究者也给出不同的改进方式.Fu 等人[13]相较于其之前工作[59]中构建的简单近邻图,进一步基于数据特征表示构建起了超图,最终在超图中进行基于随机游走的标签传播过程(该文献中,多视图的方式也可以缓解映射过程中的 Domain Shift 问题).Fu 等人[14,60]考虑嵌入空间中存在更为丰富的流形结构,使用类标签图对嵌入空间中的流形进行建模,对空间的距离度量计算采用了吸收马尔可夫链过程(absorbing Markov chain process,简称 AMP)而非传统的余弦或者欧式距离.Changpinyo 等人在文献中[87,96]提出了分类器合成模型,该模型考虑了模型空间(分类器参数空间)和语义空间的复杂流形分布,并在语义空间和模型空间分别引入一组了伪基类(phantom class)(这些伪基类能构成各种真实类),与真实类一起构建了加权图;通过马氏距离计算图中边的权重,进而计算类别之间的条件概率;最后,通过模型空间中节点对于权重图的嵌入来进行对齐的操作.测试过程中,通过训练好的模型直接合成分类器来进行不可见类的识别.Yanan 等人[79]提出了一种多模态数据流形对齐的度量方式,考虑到数据可能存在的复杂流形结构,将目标节点的 k 近邻节点取均值作为类原型的表示.

2.2.3 基于数据概率分布的模型

基于数据概率分布的模型,其本质上是要模拟数据高级特征的生成规律,主要可以实现两个目标:其一,使得模型可以通过这一规律来生成同类型的伪样本特征,从而将零样本问题转化为标准的监督学习问题;其二,可以让模型在生成规律的层次上进行不同模态数据的对齐操作,缓解模态鸿沟问题,变得更加鲁棒.下面我们根据构建模型技术的不同,将相关工作分为非生成式模型和生成式模型,分别进行详细的介绍.

- 非生成式模型

最开始,一些研究者通过非生成式的模型来建模数据分布.在这类方法下,不同研究者的思路千差万别.Mukherjee 等人[97]借助高斯词嵌入方法[185]将不同模态的数据建模为高斯分布,然后建立不同高斯分布之间映射关系.文献[30,100]中也是将数据建模为高斯分布.Micaelli 等人[186]面对许多数据集并不公开的情况,利用符合高斯分布的随机噪声来生成伪数据,并通过依次迭代的最大化和最小化学生网络(student network)和教师网络(teacher network)预测之间的 KL 散度,最终使得 Student Network 在不依靠任何数据或元数据的情况下,与

Teacher Network 的预测相匹配.该文献还提出了新的度量标准来量化教师网络与学生网络在决策边界附近的信念匹配程度.Guo 等人[98]依据可见类与不可见类之间的关系,使用线性映射的方法来估计每个不可见类的条件概率,进而生成不可见类的样本.Bucher 等人[99]基于降噪自编码器(denoising autoencoder)和对抗自编码器(adversarial autoencoder)的模型:前者与标准自编码器的区别在于在输入层增加了噪声输入,在隐层增加了类别信息的输入;后者在前者的基础上,引入对抗训练来对隐层潜在特征的生成进行约束,使编码分布与固定的先验分布相匹配.在这两个模型中,隐层生成的潜在特征编码信息可以视作是数据的分布信息.在其他类型任务中,文献[22,30]分别将这一思想用在了活动识别领域、零样本动作识别领域任务中.其中,Antol 等人[22]将这类模型运用在了活动识别领域,先将模型在草图上进行训练,然后在真实的图像上进行测试,训练过程中的草图可以视作由人类定义的、数据特征的本质分布;Mishra 等人[30]基于双向映射的思想提出了一个生成式模型框架,将语义空间映射到视觉空间然后再重构回去(这一映射可以是线性或者是非线性的),其中,视觉空间的每个类被建模为高斯分布.

- 生成式模型

更多研究者通过生成式的模型来完成零样本学习任务,其中绝大部分研究者通过 GAN 来拟合数据的特征分布,并生成伪样本.Tong 等人[105]认为:在映射过程中加入流形的知识虽然可以使得模型更加鲁棒,但是数据流形本身可能存在着的复杂结构(类分布重叠)也会极大地影响 Hinge Loss 或者回归损失的训练性能.因此,Tong 等人在建立共同映射模型的基础上,整合了生成对抗网络来生成两种类型的样本,分别用来增加同一类样本的多样性和提升存在重叠分布的类之间的判别性.Xian 等人[106]基于 WGAN(Wasserstein GAN)来构建模型,能够使得训练过程更加稳定;而且在 GMMN(generative moment matching network)模型[99]的基础上增加了生成伪样本的分类损失,这些举措都有助于提升所生成伪样本特征的判别性.该模型奠定了之后绝大多数基于 GAN 模型的基础架构.在此基础上,针对伪样本生成的质量问题,Li 等人则[83]进一步增加了灵魂样本(soul sample)的正则化项以及针对伪样本特征置信度的计算,其中,灵魂样本是指每个类所包含多个类原型,生成的伪样本只需靠近其中之一的表示即可.这样能够增加生成伪特征的多样性.Liu 等人[90]则加入类原型进行修正,来提升所生成的不同类伪样本之间的区分性.Paul 等人[108]基于 GAN 提出了一个直推式的零样本模型,其在目标域也训练了一个生成器和判别器,并增加目标域生成器与源域生成器参数相似的约束,从而缓解 ZSL 中的 Domain Shift 问题;除此之外,模型中还通过语义判别损失和语义关联损失相结合预训练了一个特征提取网络,经过该网络提取的图像特征在保持类内相似性关系的同时也能保持其判别性,从而减轻 ZSL 中的 Hubness 问题.需要指出的是,因为在目标域的不可见类中并不存在对应的图像-文本对,因此在直推式部分输入的数据是不存在对应关系的,这样的处理方式其实包含了“源域和目标域的数据服从同一个概率分布”这一隐藏的前提假设.文献[109,110]则将 GAN 与双向映射模型的思想结合了起来.在其他类型任务中,文献[32,37,44]分别将 GAN 扩展到了零样本动作识别、零样本语义分割和零样本视频分类任务中.其中,Mandal 等人[32]在条件 WGAN 中加入了数据分布检测器(out-of-distribution detector)来判别源域和目标域的动作类别;Bucher 等人[37]在零样本语义分割任务中,除了常规的生成网络结构之外,还使用了 GCN 来融合图内各个语义类别的信息,最终生成融合上下文信息的语义表示;Zhang 等人[44]通过增加多层次信息推断损失和互信息相关约束措施来最大化地保持不同模态信息的一致性,从而提升生成的伪样本质量.

有的研究者基于 VAE 来构建模型,VAE 相较于 GAN,其训练稳定性更好.其中,文献[86,101−104]利用单个 VAE 来学习数据的概率分布,文献[102,104]使用了 CVAE 模型.Yu 等人[104]将不可见类的数据视作可学习的变量,通过类似于 EM 算法的迭代学习策略,即重复生成伪数据的过程和参数学习的过程,来最终完成模型的训练.在该文献中也提到了生成的伪样本特征置信度问题(类似的提法还有文献[83]),并通过 dropout 操作来进行置信度的度量.而 Schonfeld 等人[103]将零样本问题视作多模态学习问题,通过减小视觉和语义空间中各自 VAE 隐层分布的 Wasserstein 距离,并增加交叉对齐损失约束,来实现不同空间数据概率分布的一致.在其他类型任务中,Yelamarthi 等人[39]将这一思路用于基于轮廓的图像检索领域,具体是将经过编码之后的真实图像特征和草图特征经过拼接输入到自编码器结构的网络中,其中,自编码器网络可以是 VAE 或者是对抗自编码器.

还有的学者将 VAE 和 GAN 相结合来处理 ZSL 任务.Huang 等人[111]将视觉-文本语义映射、文本语义-视觉映射以及度量学习(metric learning)方法融合在统一的框架下,分别对应到所提出模型的生成器模块、回归器模块和判别器模块.其中,判别器损失受文本生成图像工作[187]的启发,通过文本语义和视觉特征的组合构成了多种形式的伪数据,能够帮助生成更加鲁棒的跨模态对应关系.Xian 等人[112]通过 VAE 解码模块和 GAN 生成器参数共享的方式,将两种模型进行了结合,这种结合方式可视作对 GAN 的生成器的输入增加了 VAE 的约束;除了零样本学习任务之外,该工作还从可视化的角度去尝试对 ZSL 的认知过程进行解释,即利用文献[188]中 Image Caption 任务的网络输入伪特征,并经过反卷积生成的图像来生成文本,观察文本内容是否与图像视觉内容相吻合.刘欢等人[113]也是基于与文献[112]的类似思路,但是为非直推的模型.

总体上来看,基于抽象知识的模型在工作模式上主要分为两种:其一是在数据映射的过程中保持数据的流形结构,以增加数据的判别性;其二是在该层次知识的基础上对多模态数据之间进行对齐操作,然后进一步开展后续的工作.由于抽象知识要比初级知识更加接近数据的本质,因此往往取得更好的效果.这也是近两三年来比较热门的研究点.

2.3 引入外部知识的零样本模型

除了挖掘数据集本身所包含的知识之外,有研究者考虑引入外部知识来进一步帮助提升模型的性能.其主要包含了两种形式的外部知识:外部描述和外部知识库.下面我们将分别进行介绍.

2.3.1 基于引入外部描述的模型

在模型输入中引入有关于类别的外部描述,主要有两个目的:首先,外部的语义描述往往包含了更多对任务有利的信息,并且有时还可以节省人工标注的成本;其次,外部引入的数据形式更加贴近实际,其中包含的噪声也能使得最终的模型更加鲁棒.

Lei 等人[118]使用了 Wikipedia Article 作为语义空间,提取文本描述的 TF-IDF 特征和图像的 CNN 特征,并通过简单的内积形式将两者学习到一个统一的潜在表示空间.Qiao 等人[119]延续了文献[154]中的建模思想,并使用词袋模型处理 Wikipedia Article,用于替代对应类别的属性表示,这样能够减轻人工搜集语义表示的负担.区别于文献[154],考虑到引入的外部描述包含了极大的噪声,Qiao 等人[119]将映射参数分解后的结果进行了更进一步的矩阵分解,分解后的两个矩阵分别作为图像的分类器权重参数和用于抑制外部引入知识(属性)的噪声.Elhoseiny 等人[120]也是提取外部文本的 TF-IDF 特征,并最终用于细粒度的图像分类任务.Zhu 等人[107]则基于 GAN 强大的特征拟合能力,利用目标域类别的 Wikipedia Articles 描述来生成对应类别的视觉特征,并通过全连接层来过滤文本特征输入所包含的噪声.在其他类型任务中,Xu 等人[25]在构建视觉-语义映射的同时,通过增加额外数据来扩展数据集,并根据与目标域的相关性进行加权,从而提升模型的泛化性能,最终用于动作识别任务.Xu 等人[43]利用外部图像来提取情感词典以及外部语料中包含的语义关系信息,共同辅助迁移从视频中提取的深度特征,最终用于视频情感识别任务.

2.3.2 基于引入外部知识库的模型

外部知识库是目前大多数人所理解知识的狭义的概念范畴.通过引入外部大型知识库并作为任务的先验信息,主要有 3 个目的:其一是进行数据挖掘和分析;其二是帮助模型提取更好的特征表示;其三是在当前纯数据驱动的模型在遭遇瓶颈时,利用大规模知识库中的显式关系,能提升现有模型对于特定任务的性能或者减轻模型对数据的依赖,并可以在一定程度上增加模型的可解释性.下面我们分别进行介绍.

- 首先,一些研究者针对外部知识库进行了一些前沿性的探索.

Rohrbach 等人[16]使用 WordNet 同义词集的定义去挖掘属性.Rohrbach 等人[139]通过实验来分析知识库取代部分现有数据的可能性,实验证明:在零样本问题中,用知识库取代人工标注属性会导致基于属性的模型分类精度下降;但在基于分类器相似度方法中,其性能达到了人工监督的水平.而且实验也表明,在语义相似性度量(SR measures)方面,不同的知识库通常会导致不同的结果:Yahoo image search 和 Wikipedia 表现较好,而 Yahoo Web search 和 WordNet 表现欠佳.Zeynep 等人[182]鉴于属性知识获取代价较大,探讨了对外部层次化的知识库或者外部描述进行编码,并辅助或者取代属性知识的可能性.Gan 等人[24]通过实验分析得出:相较于基于 WordNet 关系

计算的词的相似度(JC 距离),基于类名词嵌入之间余弦相似度构建的可见类与不可见类之间的关系,能够更好地进行知识传播过程(但仅限于动作识别领域).Kordumova 等人[36]通过引入外部信息以及知识库(WordNet)去识别图像中的场景,并不使用任何场景图像作为训练数据,并通过实验分析得出:来自知识库中间层次的目标对于场景的识别有较大贡献,而分别来自顶层和底层的 General 目标和 Fine-Grained 目标则对场景识别贡献有限.

有的研究者受限于现有知识库对于某些任务的局限性,根据特定任务特性自己定义知识库,从而更好地完成相应任务.Deng 等人[34]鉴于多分类问题中标签相互独立的假设并不成立的问题,自己定义了一个 HEX (hierarchy and exclusion)图,图中的语义关系可以分为 Mutual Exclusion、Overlap 和 Subsumption,然后以该图作为标签关系先验,构建了一个基于条件随机场的概率分类模型.

- 其次,有的研究者引入知识图谱,并侧重于使用其层次化的知识表示形式.

Al-Halah 等人[126]利用了知识库中层次化的分类信息,在不同层次上进行属性学习,并进行层次化的属性迁移.Li 等人[127]沿用了文献[163]中学习映射的思想,借助 WordNet 知识库进行了层次化的文本语义嵌入,并假定每个标签在 WordNet 中均有对应节点,从根节点到特定节点,越靠近特定节点的节点,其贡献越大,最后将这一思想加入到凸化组合中.Li 等人[128]则利用了知识库中层次化的分类信息,用于提取更加具有判别性的图像特征,然后基于提取的源域图像特征进行域适应和标签传播操作,最终进行细粒度的类别认知.DUTTA 等人[40]将自编码器和 GAN 结合,利用知识库(WordNet)中层次化的表示并结合词嵌入来引导自编码器生成更加具有判别性的特征表示,从而更好地辅助 GAN 进行对抗的训练,最终进行基于轮廓的图像检索任务.

- 最后,有研究者将“符号主义”和“连结主义”中的表示方法相结合,成为当前的主流形式.

本文根据对符号知识表示方法的应用,又分为传统方式和基于深度学习(如图网络)的方式.

有研究者将知识图谱用传统的知识表示方法(如翻译模型)进行表示.这一类模型更多地出现在传统图像分类任务之外.Lu 等人[46]将传统的知识表示与零样本视觉关系检测任务相结合,所构建的模型由两个模块组成.

1) Visual Appearance Module:训练 VGG 网络用于提取图像中的 Object 和 Predicate.
2) Language Module:将两个 Object 拼接为新的向量来表示视觉关系三元组,然后通过映射函数来使得三元组之间的关系正比于它们所包含的 Predicate 对应词嵌入之间的余弦距离,其值越大,表示对应的视觉三元组成立的概率就越高.

最后做 Triplet Loss(rank loss),并且区别于之前的数据集只包含较少的视觉关系类型,Lu 等人创建了一个新的数据集 VRD,包含了数万种关系.Cui 等人[129]将传统知识表示用于零样本图像的多标签分类任务中,所提出模型将知识(ConceptNet 知识库)表示与多标签的图像表示结合在一起,两者进行协同的训练,即将图像分类分支的分类器权重参数与知识表示分支的节点映射参数进行了共享,在完成图像多标签分类的同时,能够实现知识库中有关系的节点表示尽可能接近.最终,在标签预测任务、零样本标签推测任务以及基于内容的图像检索任务中证明了模型的有效性.并且实验表明:该模型可以在某种程度上提炼知识库来描述图像,并使用结构化标记来标记图像.

有研究者将现有任务模型与图网络相结合.Wang 等人[130]首先将图网络与知识库结合并用于零样本图像分类任务中.模型分为两个独立的部分:CNN 分支和 GCN 分支.CNN 分支首先使用预训练好的 CNN 网络为原始图像抽取高级特征;其次,GCN 分支(如图 6 所示下方的 GCN 网络,模型示意图来源:https://github.com/JudyYe/zero-shot-gcn)将数据集中的每个类别作为知识图中的一个节点,并对其词嵌入表示作为节点的初始输入.模型训练时,可见类节点的初始表示经过 GCN 网络的信息融合,融入了周围节点的信息并形成新的表示,然后知识图中可见类节点范围内,利用来自 CNN 部分的图像类别对应的分类器权重参数作为监督信息(图 6 所示绿色节点)来训练 GCN 模型的参数.测试时,将知识图中的不可见类节点的输出视作对应类别的分类器权重参数.需要指出的是,该模型使用的知识图谱是基于 NELL[189]和 NEIL[190]构造的新知识图谱.Kampffmeyer 等人[131]则对前面的模型[130]进行了进一步改进,包括:(1) 使用了更少的图卷积网络层数来避免训练过程中节点表示的趋同性;(2) 进一步地改进现有 WordNet 知识库,使其节点之间的连接更加密集,并根据节点间的距离加入了注意力机制(attention mechanism,简称 AM);(3) 在训练过程中采用轮流优化策略,固定 GCN 的参数,对预训练好的 CNN

进行微调操作来缓解 Domain Shift 问题.这些操作均进一步的提升了模型效果.

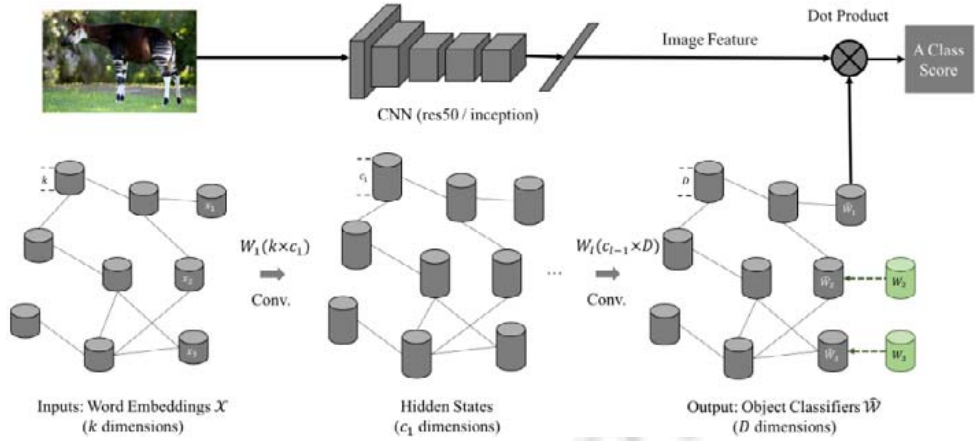


Fig.6 Architecture of GCNZ (GCN for zero-shot learning)[130]

图 6 GCNZ(GCN for zero-shot learning)框架[130]

Zhang 等人[62]指出了之前基于图网络模型[130,131]的不足:首先,它们仅基于可见类构建损失进行训练,而不涉及到不可见类,因此域偏移问题仍然存在;其次,该关系仅在类级别层次建模,忽略了实例级关系,导致数据的判别能力不足;最后,这些方法对关系的利用仍然是隐式的(指不是直接利用关系进行标签传播,而是借助关系将节点表示转化为分类器参数),这会导致在最终分类的过程中,被提炼出来的知识被稀释.针对前面这些问题,Zhang 等人提出了 TGG(transferable graph generation)模型.

TGG 由两个模块组成——GraphGeneration 和 RelationPropagation.

- 在 GraphGeneration 阶段,首先构建了 Class-Level Prototype 图,该图是借助 ConceptNet 知识库包含的显式关系进行构建的,各个节点的表示为视觉特征,不可见类的节点由 GAN 生成的伪样本作为输入;在此基础上进行 Multi-Head Attention+Multi-Level Attention 机制的训练,修正节点表示,使得数据更加具有判别性;最终经过关系核(relation kernel)损失(即生成的新图需要和对应的原图结构保持一致,防止第 1 阶段训练过拟合)生成了 Instance-Level 图,进入到 RelationPropagation 阶段.
- 在 RelationPropagation 阶段,使用标签传播算法(相比隐式嵌入方法,这样能使知识的传播更有效率),并构建双向传播机制(分别将图的可见类部分和不可见类部分作为初始标签矩阵),最终使用元学习的训练策略来训练模型.

需要指出的是,注意力机制、双向标签传播以及元学习的训练机制均是用来缓解域偏移问题的,而该模型可以用于完成 ZSL,GZSL 以及 FSL 任务,其框架示意图如图 7 所示($\mathcal{G}_c$ 表示类级别的图,$\mathcal{G}_l$ 表示样本级别的图).

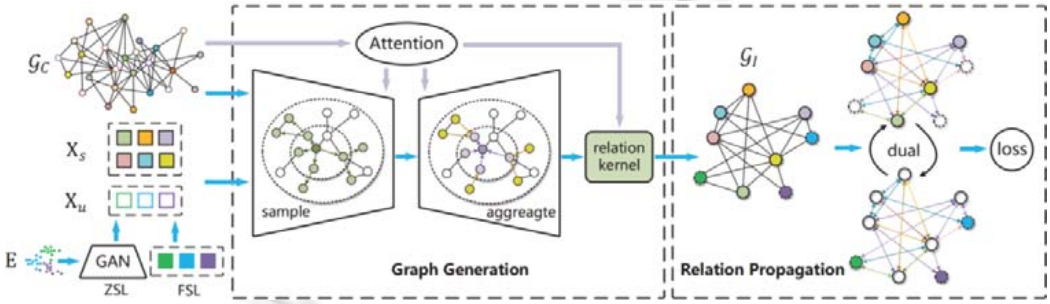


Fig.7 Architecture of TGG[62]

图 7 TGG[62]框架

除了传统的图像分类任务之外,Lee 等人[132]将图网络扩展到零样本多标签图像分类领域,利用 WordNet 知识库中的显式关系为图像标签构建图关系表示,然后基于标签词嵌入向量之间的相似度来建模关系的权重,最后将图像特征和标签表示作为初始的节点状态,使用图门神经网络(graph gated neural networks,简称 GSNN)[191,192]来进行知识传递.Zhang 等人[31]进一步将 ConceptNet 5.5 知识库引入到动作识别领域,设计了一个两分支的图卷积网络:一个分支用于生成分类器参数,另一个分支用于生成实例,从而有效地对动作-属性、属性-属性以及动作-动作之间的关系进行建模.

总体来看,引入外部知识的模型,其作用更多体现在通过增加人类的知识先验来进一步减小模型对当前数据的依赖,并提升特定任务的性能.但是这同时也意味着需要对外部知识进行噪声处理,以尽可能消除对任务不利的影响.

3 存在的问题和模型总结

3.1 存在问题

在视觉领域的零样本学习任务中,长期以来存在两个传统的问题:域偏移问题和枢纽点问题,下面分别进行详细的介绍.

- 域偏移问题

即 Domain Shift 问题,该问题由 Fu 等人[13]首次提出.该文献中将问题定义为“由于源域数据集和目标域数据集包含不同的类,因此这些类所包含数据分布也可能不同.在源域数据集上学习的从视觉空间到嵌入空间的映射函数,如果没有针对目标域数据集进行任何的调整,就会产生未知的偏移/偏差”.在 Fu 等人[14]的工作中,对这一问题进行了扩展,由原先源域到目标域(projection domain shift)映射偏移的问题扩展到不同模态映射偏移(visual-semantic domain shift)的问题.从本质上看,前者可以简单理解为域适应问题,后者可以简单理解为模态鸿沟(又被称为语义间隔)问题.

针对第 1 种类型的域偏移问题,Long 等人[88]阐述为“这个问题是因为基于回归形式的模型无法发现语义空间固有的几何结构,也不能捕捉到可见类到不可见类之间的关系”.因此针对这个问题,最好的解决方式就是在训练过程中融入不可见类(目标域)的信息(利用不可见类的流形信息),使得模型更好地捕获源域与目标域之间的关系,从而增加模型在目标域的域适应性.由这一思想产生了两种主流的做法:第 1 种是建立直推式的模型[10−13,25,28,30,31,42,43,53,59,77,79,85,93,94,100,101,108,112,153,173−175,177],即在训练过程中加入不可见类的样本;第 2 种是通过生成伪样本(基于 GAN[32,40,44,83,90,99,105−113]、非 GAN[67,88,98,102,104,145,151]),将零样本问题转化为标准的监督学习问题.这些做法其实均隐含了一个前提假设条件:“目标域与源域的数据分布在样本级别上是一致的”.Liu 等人[70]则放弃寻求样本级别的一致性,转而寻求任务级别上的一致性.Wan 等人[77]则直接利用目标域的不可见类样本进行 *k*-means 聚类来获取目标域的数据分布.

此外,除了在训练过程中加入目标域的信息的方式之外,也有研究者通过保留源数据足够多的信息来缓解 Projection Domain Shift 问题.主要有两种方式:一种是建立双向映射[66,67,101,109,160−162],经过特征空间的重构从而建立更加鲁棒的映射模型;另一种是通过增加流形正则化项来保持数据的结构[28,29,88,92,93,161].最后,还有 Zhang 等人[62]在基于所生成伪数据的基础上,通过引入元学习的训练机制来减轻 Projection Domain Shift 问题.

针对第 2 种类型的域偏移问题,比较典型的处理方式是基于数据的抽象知识去构建模型,并在抽象层次上进行不同模态数据的对齐操作.有以下的两种方式:第 1 种是利用流形对齐的思路,从数据分布的本质特征角度出发,去进行多模态空间的流形对齐[53,79,80,87,94,96],但是 Wang 等人[151]对“不同空间中的数据分布一致”这一假设过于严格的问题进行了处理;第 2 种是从数据概率分布的角度出发,将多模态数据的特征概率分布进行对齐[30,86,97,100,101,103].

- 枢纽点问题

即 Hubness 问题,其可以阐述为“在特征空间中,某个点会成为大多数节点的最近邻点(即使它们之间无关),这会导致数据失去其判别性”,尤其会影响基于最近邻的零样本认知方式的最终效果.Dinu 等人[15]通过实验发

现了 Hubness 问题的存在,并将 Hubness 问题阐述为“Hubness 问题是高维空间的固有问题”,会极大地影响基于回归映射的方法.Lazaridou 等人[155]将 Hubness 问题具体阐述为“高维空间经常受到中心性(hubness)的影响,也就是说,它们包含某些元素(即中心点),这些元素会靠近空间中的许多其他的点,却并不与后者相似”.这些论述均表明,Hubness 问题是在高维空间中的一个固有现象.接着,在 Shigeto 等人[158]的工作中,通过实验的分析表明:Hubness 问题的出现不仅仅是因为高维空间,而且和 Ridge Regression 岭回归方法在零样本问题中的使用方式有关.作者还进一步讨论了基于 Ridge Regression 的模型受到的 Hubness 问题的影响.

对于 Hubness 问题,有 3 种主流的解决方式.

1) 其一是更新映射的方式.基于文献[155]的论述,很多研究者[11,68,77,89,158]革新了映射的方式,建立了反向的映射(从文本到视觉,从低维到高维)来减轻 Hubness 问题对结果的影响;但在文献[193]中证明了,Hubness 问题在低维空间中也会存在,因此这一解决方式并不彻底.
2) 其二是基于原有模型增加流形正则化项.从上面的论述中可以看出,Hubness 问题更多是基于回归映射的模型所存在的本质问题,即映射会导致数据的判别性降低.因此,很多研究者[28,29,88,92,93,161]通过增加流形正则化项来保持数据的流形结构,进而保持数据的判别性.
3) 其三是转换建模思路.有的研究者不使用基于回归映射的模型,转而通过生成伪样本,将零样本类的认知过程转化为一个标准的监督学习问题,从而也避免了 Hubness 问题对结果的影响.但是伪样本也需要进行筛选以保证质量.

此外还有一些非主流的方式,例如,Dinu 等人[15]提出一种基于全局修正的近邻搜索方法而非最近邻搜索的零样本认知形式;在文献[155]中,Lazaridou 等人将岭回损失替换为 Max-Margin Ranking Loss 来缓解 Hubness 问题.

3.2 模型总结

本文从数据知识的角度出发,依据知识的来源途径将知识的定义划分为“初级知识、抽象知识和外部知识”,并基于这样的划分方式将现有相关工作分为“基于初级知识的模型、基于抽象知识的模型以及引入外部知识的模型”.在每部分内容中,我们基于模型对该层次知识的利用方式,分别对其进行了梳理和归纳总结.更重要的是,这样的架构也有助于我们理解模型逐渐克服 ZSL 中存在的各种问题的过程.基于本文第 3.1 节针对问题的论述,我们对 3 类模型进行了总结,并将它们总体表现出的优缺点呈现在表 2 中,以便更好地看出 ZSL 技术发展的脉络和趋势.

Table 2 Comparison of advantages and disadvantages in different zero-shot learning methods

表 2 零样本学习各类方法优缺点对比

模型分类		优点	缺点
基于初级知识的模型	概率模型	1. 人工定义的属性具有明确的描述性,解释性较好	1. 属性的获取代价较大; 2. 人工定义的属性维度较小,且每个维度描述内容之间并不是完全独立的; 3. 人工定义的属性表示与视觉表示之间存在鸿沟(模态鸿沟); 4. 自动学出的属性表示没有很好的描述性; 5. 难以扩展到增量学习的范畴
	基于映射的模型	1. 基于映射的模型较为直接地构建了视觉空间与语义空间之间的关联,直观易于理解; 2. 基于反向映射的模型能够缓解 Hubness 问题对模型结果带来的影响; 3. 基于双向映射的模型能够保留更多的数据信息,缓解 Domain Shift 问题	1. 基于映射的模型无法很好地反映类内方差这一特性; 2. 词嵌入表示相较于人工定义的属性,与视觉表示之间存在更严重的模态鸿沟问题; 3. 绝大多数模型均基于 k 近邻分类器进行识别,没有高效的判别分类器. 4. 基于回归映射的模型比较容易受到 Hubness 问题的影响; 5. 模型及实验结果解释性较弱

Table 2 Comparison of advantages and disadvantages in different zero-shot learning methods (Continued)

表 2 零样本学习各类模型优缺点对比(续)

模型分类		优点	缺点
基于抽象知识的模型	基于类原型的模型	1. 相较于语义空间的类原型表示,基于视觉空间的多类原型表示能够在一定程度上反映出类内方差; 2. 在视觉类原型基础上建模(基于映射的模型),相较于基于样本个体建模,其计算效率更高; 3. 视觉空间的类原型表示可以更好地关联到视觉分类器参数特征	1. 对数据进行视觉类原型的表示,会损失部分数据的分布信息
	基于数据流形分布的模型	1. 流形正则化项能够有效保持数据的结构,保留更多的数据信息,从而缓解 Domain shift 问题和 Hubness 问题; 2. 基于数据流形分布特征的多模态数据之间的对齐,能够更好地缓解 Domain shift 问题	1. 数据可能存在复杂的流形分布,难以正确地模拟其分布; 2. 源域和目标域的数据流形可能存在不一致的情况,难以合理地进行建模
	基于数据概率分布的模型	1. 从数据概率特征分布的层面进行多模态对齐,能够更好地缓解 Domain Shift 问题; 2. 生成伪样本的过程,将零样本学习问题转化为标准的监督学习问题,避免了 Hubness 问题,也能在一定程度上缓解 Domain shift 问题	1. 生成伪样本的质量及其多样性有待提高,并难以进一步扩展到大规模数据集上(尤其是基于 GAN 的模型); 2. GAN 的训练稳定性问题; 3. 源域和目标域的数据概率分布可能存在不一致的情况,难以进行合理的建模; 4. 模型及实验结果解释性较弱
引入外部知识的模型	引入外部描述的模型	1. 扩充了数据来源; 2. 能够训练出更加鲁棒的模型	1. 外部描述包含的噪声较多,完全过滤掉对任务不利的噪声手段有限; 2. 模型及实验结果解释性较弱
	引入外部知识库的模型	1. 基于外部知识库构建的模型(尤其是基于图网络的模型),性能较好,模型得出结果的解释性相较于一般的深度网络更好; 2. 能够比较容易地扩展到大规模数据集上	1. 图神经网络的训练会出现节点趋同(过拟合)的现象; 2. 知识的表示形式仍然不够合理(多义性)

4 数据集、评估标准和实验

由于在零样本学习领域,图像分类任务是主流,因此,本节将介绍零样本图像分类任务中的常用数据集,并且基于当前“数据+知识驱动”的背景,进一步介绍了基于外部知识库的模型中常用的知识图谱.最后,还介绍了 ZSL 和 GZSL 两个分类任务的评估标准.

4.1 常用数据集

绝大多数零样本图像分类模型所用的数据集包含了 AWA(animal with attribute)数据集[33]、AWA2(animal with attribute 2)数据集[4]、CUB(Caltech-UCSD Birds-200-2011)数据集[194]、SUN(SUN attributes)数据集[195]、FLO (Oxford 102 flowers)数据集[196]和 aPY(aPascal-aYahoo)数据集[134].上述 6 个数据集属性见表 3.

Table 3 Introduction of datasets (image classification) properties

表 3 数据集(图像分类任务领域)属性介绍

数据集	AWA	AWA2	CUB	SUN	FLO	aPY
图像样本数	30 475	37 322	11 788	14 340	8 189	12695+2644
类别(训练集/测试集)	40/10	40/10	150/50	645/72	82/20	20/12
属性维度	85	85	312	102		64
属性值(实质或布尔值)	兼有	兼有	兼有	布尔		兼有

需要指出的是,AWA2 数据集是 AWA 数据集版权到期之后该数据集的替代;CUB 数据集中的每幅图像都用 Bounding Boxes 和 Part Locations 进行了标注,并被用于细粒度的图像分类任务;SUN 数据集是用于细粒度场景分类的 SUN 数据库[197]的一个子集;在 FLO 数据集中,不同的研究者给出了每个类别不同的对应文本语义描述[166,198];aPY 数据集包含来自于 PASCAL VOC 2008 数据集的 20 个类别以及来自于 Yahoo 的 12 个类别.

除上述 6 个通用的数据集外,ImageNet 数据集[199]也是目前零样本图像分类任务领域越来越广泛使用的大

规模数据集.该数据集根据 WordNet 的层次结构进行组织,因此 ImageNet 数据集中的所有类都能在 WordNet 中找到对应节点.完整的 ImageNet 数据集包含了大约 22 000 个类别,超过 1 500 万张标签高分辨率图像,由 Amazon s Mechanical Turk (AMT)众包工具进行标记,被称为 ImageNet 21k 数据集.该数据集存在较大的类别不均衡问题,因而是当前同类任务中最具挑战性的数据集.Xian 等人[4]的工作中,总结了前面具有代表性的方法在该数据集上的实验效果;最近的一些方法也同样在 ImageNet 21k 数据集上进行了验证[4,103,130,131].其使用情况大致如下:首先,使用 ImageNet 1k 进行模型的训练;然后,测试集分为 3 个级别——2-hop,3-hop 和 all,其中,2-hop 和 3-hop 分别是指在 WordNet 中,距离 ImageNet 1k 类对应节点 2-hop/3-hop 距离的节点所对应类作为测试类,all 则代表了剩余的 20k 的类别;除此之外,还有模型使用除 ImageNet 1k 之外的剩余类别中最受欢迎的 500/1k/5k 等类别,以及最不受欢迎的 500/1k/5k 的类别进行测试.但在基于生成式模型的方法中(尤其是指基于 GAN 的模型),由于其生成的伪样本质量不能得到充分的保证,因此向 ImageNet 21k 这种大规模的数据集扩展仍具有较大困难.

由于 ImageNet 21k 过于庞大,因此进一步衍生出了 ImageNet 1k 数据集(ILSVRC),其包含 1 000 个类别,每个类别大约有 1 000 张图片.有的研究者使用该数据集来测试模型性能,例如,Yanan 等人[79]将 ILSVRC 2012 数据分为 800/200 类用于训练/测试;文献[61,86,102]则以 ILSVRC 2012 的训练集为源域数据,并以 ILSVRC 2012 的测试部分和 ILSVRC 2010 的数据(或者不与 ILSVRC 2012 重合的 ILSVRC 2010 类别)作为目标域数据等.但是显然,ImageNet 21k 是未来工作的主流.

4.2 常用知识库

- WordNet(知识图谱发展报告 2018)

WordNet 是最著名的词典知识库,主要用于词义消歧,其表示框架主要定义了名词、动词、形容词和副词之间的语义关系,例如名词之间的上下位关系(如“猫科动物”是“猫”的上位词)、动词之间的蕴含关系(如“打鼾”蕴含着“睡眠”)等.在 WordNet3.0 中,已经包含超过 15 万个词和 20 万个语义关系.在零样本任务领域,主要使用的是 WordNet 知识库中的名词部分.在这部分内容中,有别于通常意义上的字典,WordNet 知识库根据词条的意义将其分组,每一个具有相同意义的字条组称为一个 Synset(同义词集合),WordNet 为每一个 Synset 提供了简短、概要的定义,并记录不同 Synset 之间的语义关系.这些语义关系通过一个层次化树状结构组织起来,并且图中节点之间的距离(JC 距离)大致可以反映出视觉上的相似性程度[133].由于 WordNet 与 ImageNet 数据集的紧密关系,WordNet 知识库成为视觉任务,尤其是图像分类任务领域的常用知识库.

- ConceptNet[121,122]

ConceptNet 是常识知识库,是具有代表性大规模网络知识获取的工作,最早源于 MIT 媒体实验室的 Open Mind CommonSense(OMCS)项目.ConceptNet 知识库以三元组形式的关系型知识构成,比较侧重于词与词之间的关系.从这个角度看,ConceptNe 更加接近 WordNet,但是又比 WordNet 所包含的关系类型多.ConceptNet5 的知识表示框架主要包含如下要素:概念 Concepts、词 Words、短语 Phrases、断言 Assertions、关系 Relations、边 Edges.Concepts 由 Words 或 Phrases 组成,构成了图谱中的节点.与其他知识图谱的节点不同,这些 Concepts 通常是从自然语言文本中提取出来的,更加接近于自然语言描述,而不是形式化的命名.Assertions 描述了 Concepts 之间的关系,类似于 RDF 中的 Statements.Edges 类似于 RDF 中的 Property.ConceptNet5.5 中已经包含了超过 2 100 万个关系描述和 800 万个节点(英语部分包含了大约 150 万个节点),其中包含了 21 个预定义的、多语言通用的关系(如 IsA、UsedFor 等)和从自然语言文本中抽取的更加接近于自然语言描述的非形式化的关系(如 on top of,caused by 等).在文献[31,129]中,研究者选取其英文表述的概念,并且 NUSWIDE 数据集和 ConceptNet 之间存在 92 595 个共享标签(包含 words 和 phrases 在内),因此也能较方便地用于视觉任务.

- NeLL[189]

该知识库由卡内基梅隆大学开发,是具有代表性的大规模网络知识获取的工作.和 ConceptNet 类似,也是遵循 RDF 数据模型的形式.其已经抽取了大约 170 万种物体实体、240 万条边.文献[130]中,将其和 NEIL[190](包含超过 1 700 条关系和超过 40 万的视觉个体)一起,构建了新的知识图谱来进行零样本认知的任务,

以上介绍的知识库包含了海量的各种人类先验知识,但是对于特定任务而言,任务不相关信息属于噪声.因

此,研究者在构建基于这类知识库的模型时,通常需要根据具体任务来对初始的知识库进行适当的筛选和改造.

4.3 常用任务评估标准

在单标签的零样本图像分类任务中,通常使用 Top-1 准确率来进行模型性能的度量.Top-1 准确率的定义为:预测概率最大的标签与真实标签相符的准确率(即每个测试类中正确标记的实例的比例).由于测试涉及到多个类的 Top-1 准确率,因此要进一步对所有的测试类求平均精确值.其公式定义如下:

$$acc_y = \frac{1}{|y|}\sum_{c=1}^{|y|}\frac{\text{第}c\text{类中正确预测的样本}}{\text{第}c\text{类中的样本数}} \tag{10}$$

其中,$|y|$表示类别数.在传统的 ZSL 设定下,$|y|$的范畴仅包含目标域的类别,但在 GZSL 的设定下,$|y|$的范畴进一步包含了源域的类别.因此在这种设定条件下,通过计算源域类和目标域类的 Top-1 精确度的调和平均值(该均值更加强调较小的一方的重要性,因为模型最终需要在源域和目标域均取得较高的精确值)来进行模型性能的度量,其公式表示如下:

$$H = \frac{2acc_{y^{tr}} \cdot acc_{y^{te}}}{acc_{y^{tr}} + acc_{y^{te}}} \tag{11}$$

其中, $acc_{y^{tr}}$ 表示源域的平均 Top-1 精确值, $acc_{y^{te}}$ 表示目标域的平均 Top-1 精确值.

在多标签的零样本图像分类任务中[132],使用 Precision(P),Recall(R)和 $F1$-*measure* 来进行模型性能的度量.接下来,通过一个例子来说明这 3 个性能度量标准的含义.假定模型通过预测,给出了某个图像的最终预测标签集合,其中有 *TP*(true positive)个确实为图像的标签并被正确判定;有 *FN*(false negative)个确实为图像的标签,但没有被正确判定,即在预测集合中没有出现;有 *FP*(false positive)个不属于图像的标签,但被错误判定为其标签,即出现在预测集合中;最后有 *TN*(true negative)个本来不属于图像的标签,也没有出现在预测标签集合中.基于上面的统计,这张图像对应的精确率(precision)和召回率(recall)的计算方式如下:

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{TP + FN} \tag{13}$$

从上面的两个计算结果中我们可以看出:精确率度量的是给出的标签预测中有多少是正确的,召回率度量的是模型正确召回了多少个正例.接下来,基于与 GZSL 任务中调和平均数相同的考虑,计算精确率和召回率的调和平均数来得出模型的 $F1$ 均值,其公式表示如下:

$$F1 = \frac{2 \cdot R \cdot P}{R + P} \tag{14}$$

这即为多标签图像分类任务的评估标准.

4.4 实 验

本节在前面第 2 节的模型梳理工作基础上,并结合文献[4]中实验部分的工作,从每个类别的方法中分别抽取了 1~2 个较新的模型来展示其实验结果,并在部分研究者们公布的源代码(COSMO[156],https://github.com/yuvalatzmon/COSMO;RKT[151],https://github.com/LiangjunFeng/Implement-of-ZSL-algorithms;文献[68],https://github.com/lzrobots/DeepEmbeddingModel_ZSL;EXEM[89],https://github.com/pujols/Zero-shot-learning-journal;BMVSc[77],https://github.com/raywzy/VSC;ListGAN[83],https://github.com/lijin118/LisGAN;ADGPM[131],https://github.com/cyvius96/DGP;CADA-VAE[103],https://github.com/edgarschnfld/CADA-VAE-PyTorch;GCNZ[131],https://github.com/JudyYe/zero-shot-gcn)基础上对相关模型进行了验证(未公布源码的模型均根据作者文中描述进行实现),算法运行平台为 GPU TITAN Xp×2,显存为 12×2GB.在表 4 中,从上到下的模型类别依次为基于属性迁移、正向映射、反向映射、双向映射、共同映射、其他映射、视觉类原型、数据流形分布、数据概率分布、引入外部描述、引入外部知识库的模型,依次对应了表 4 中第 1、2-3、4-5、6、7、8-10、11、12、13-18、19、20

个模型.表 5 则为对应模型在 GZSL 中的实验结果.需要指出的是,表格中带*号的是直推式的模型,字体加粗的模型则打破了“样本级别,源域目标域数据分布一致”潜在假设.需要注意的是,表 4 和表 5 中‘SS’和‘PS’的定义与文献[4]保持一致,分别表示传统的数据集分割标准和新提出的数据集分割标准.后者在一定程度上防止了预训练增益,使得在该标准下的实验结果更具科学性.而表 5 中 ts、tr、H 分别表示模型在 GZSL 任务中,目标域类别、源域类别的实验效果以及前两者的调和平均数.

Table 4 Traditional ZSL experiments of different models on various datasets, measuring top-1 accuracy (%)

表 4 在不同数据集中,各个模型在传统 ZSL 任务中的 Top-1 准确率(%)

模型	AWA		AWA2		CUB		SUN		FLO		aPY	
	SS	PS	SS	PS	SS	PS	SS	PS	SS	PS	SS	PS
DAP[33]	57.1	44.1	58.7	46.1	37.5	40.0	38.9	39.9	–	–	35.2	33.8
COSMO[156]	–	55.9	–	–	–	35.6	–	21.0	–	58.1	–	–
RKT[151]	**81.31**	–	–	–	**46.24**	–	–	–	–	–	–	–
文献[68]	88.1	–	–	–	59.0	–	–	–	–	–	–	–
EXEM[89]	76.5	–	–	–	58.5	–	67.3	–	–	–	–	–
DLFZRL[162]	–	66.3	–	63.7	–	57.8	–	59.3	–	–	–	44.5
PREN[10]	–	–	95.7	74.1	66.9	66.4	63.3	62.9	–	–	–	–
TCN[179]	–	70.3	–	71.2	–	59.5	–	61.5	–	–	–	38.9
DCN[69]	82.3	65.2	–	–	55.6	56.2	67.4	61.8	–	–	–	43.6
CPL[70]	–	–	–	**72.7**	–	**56.4**	–	**62.2**	–	–	–	**45.3**
BMVSc*[77]	**95.9**	–	**96.8**	**81.7**	**73.6**	**71.0**	**66.2**	**62.2**	–	–	–	–
CDL[53]	–	69.9	–	–	–	54.5	–	63.6	–	–	–	43.0
CVAE[102]	–	71.4	–	65.8	–	52.1	–	61.7	–	–	–	
ListGAN[83]	–	70.6	–	–	–	58.8	–	61.7	–	69.6	–	43.1
SABR-I[108]	–	–	–	65.2	–	63.9	–	62.8	–	–	–	–
SABR-T*[108]	–	–	–	88.9	–	74.0	–	67.5	–	–	–	–
f-VAEGAN-D2[112]	–	70.3	–	–	–	72.9	–	65.6	–	70.4	–	–
f-VAEGAN-D2*[112]	–	89.3	–	–	–	82.6	–	72.6	–	95.4	–	–
文献[119]	–	66.46	–	–	–	29	–	–	–	–	–	–
ADGPM[131]	–	–	–	74.6	–	–	–	–	–	–	–	–

Table 5 GZSL experiments of different models on various datasets, measuring top-1 accuracy (%)

表 5 在不同数据集中,各个模型在 GZSL 任务中的 Top-1 准确率(%)

模型	AWA			AWA2			CUB		
	ts	tr	H	ts	tr	H	ts	tr	H
DAP[33]	0.0	88.7	0.0	0.0	84.7	0.0	1.7	67.9	3.3
COSMO[156]	64.8	51.7	57.5	–	–	–	41.0	60.5	48.9
EXEM[200]	–	–	58.3	–	–	–	–	–	35.6
DLFZRL[162]	–	–	40.5	–	–	45.1	–	–	37.1
PREN[10]	–	–	–	32.4	88.6	47.4	35.2	55.8	43.1
TCN[179]	49.4	76.5	60.0	61.2	65.8	63.4	52.6	52.0	52.3
DCN[69]	25.5	84.2	39.1	–	–	–	28.4	60.7	38.7
CPL[70]	–	–	–	**51.0**	**83.1**	**63.2**	**28.0**	**58.6**	**37.9**
BMVSc*[77]	–	–	–	**71.9**	**88.2**	**79.2**	**33.1**	**86.1**	**47.9**
CDL[53]	28.1	73.5	40.6	–	–	–	23.5	55.2	32.9
CVAE[102]	–	–	47.2	–	–	51.2	–	–	34.5
CADA-VAE [103]	72.8	57.3	64.1	75.0	55.8	63.9	53.5	51.6	52.4
ListGAN[83]	52.6	76.3	62.3	–	–	–	46.5	57.9	51.6
SABR-I[108]	–	–	–	30.3	93.3	46.9	55.0	58.7	56.8
SABR-T*[108]	–	–	–	79.7	91.0	85.0	67.2	73.7	70.3
f-VAEGAN-D2[112]	57.1	76.1	65.2	–	–	–	63.2	75.6	68.9
f-VAEGAN-D2*[112]	86.3	88.7	87.5	–	–	–	73.8	81.4	77.3
DAP[33]	4.2	25.1	7.2				4.8	78.3	9.0
COSMO[156]	35.3	40.2	37.6	59.6	81.4	68.8	–	–	–
EXEM[200]	–	–	–	–	–	–	–	–	–
DLFZRL[162]	–	–	24.6	–	–	–	–	–	31.0
PREN[10]	35.4	27.2	30.8	–	–	–	–	–	–

Table 5 GZSL experiments of different models on various datasets, measuring top-1 accuracy (%) (Continued)

表 5 在不同数据集中,各个模型在 GZSL 任务中的 Top-1 准确率(%)(续)

模型	SUN			FLO			aPY		
	ts	tr	H	ts	tr	H	ts	tr	H
TCN[179]	31.2	37.3	34.0	–	–	–	24.1	64.0	35.1
DCN[69]	25.5	37.0	30.2	–	–	–	14.2	75.0	23.9
CPL[70]	**29.1**	**32.4**	**26.1**	–	–	–	**19.6**	**73.2**	**30.9**
BMVSc*[77]	**29.9**	**62.9**	**40.6**	–	–	–	–	–	–
CDL[53]	21.5	34.7	26.5	–	–	–	19.8	48.6	28.1
CVAE[102]	–	–	26.7	–	–	–	–	–	–
CADA-VAE[103]	35.7	47.2	40.6	–	–	–	–	–	–
ListGAN[83]	42.9	37.8	40.2	57.7	83.8	68.3	34.3	68.2	45.7
SABR-I[108]	50.7	35.1	41.5	–	–	–	–	–	–
SABR-T*[108]	58.8	41.5	48.6	–	–	–	–	–	–
f-VAEGAN-D2[112]	50.1	37.8	43.1	63.3	92.4	75.1	–	–	–
f-VAEGAN-D2*[112]	54.2	41.8	47.2	91.0	97.4	94.1	–	–	–

总体上来看:从初级知识到抽象知识的发展过程中,各类模型的识别准确率是不断上升的.这是因为抽象层次的知识相比于初级知识更加接近数据分布的本质.以常用的 AWA 数据集为例,通过分析表 4 和表 5,我们能够印证本文第 3 节得出的一些结论.

(1) 从模型本身结构的角度来看,在基于初级知识的模型中,基于映射的模型是已有方法的主流.针对领域内问题从而对映射方式所做的改进,使得模型效果也在不断提升.从表 4 中我们可以看出,反向映射、双向映射和共同映射以及其他方式的映射确实能对模型效果提升带来较大促进作用,因为他们相对于正向映射建立了更加鲁棒的映射关系.这一点我们从表 4 的 ZSL 任务中(从第 2 类方法到第 6 类方法)可以明显看出:相对于正向映射,平均提升效果 20.3%.而在基于抽象知识的模型中,依赖于生成式模型的强大拟合能力,能够挖掘出数据的内在分布规律,因此,基于数据概率分布的模型普遍取得了更好的实验效果.相较于基于映射的模型,平均提升效果 17.0%.从表 4 和表 5 中可以看出,基于 GAN 的模型相比于基于 VAE 的模型普遍效果更好(例如 ListGAN 和 CVAE 相比,提升效果 31.9%).这是因为基于 GAN 的模型生成伪样本的能力更强,这类方法将成为今后的主流.

(2) 从数据利用的角度来看,在训练过程中融入不可见类的数据,即将模型由归纳式改造为直推式,往往是能够提升模型识别不可见类精度的最简单有效的方法,这从表 4 同一个模型(例如 SABR 和 f-VAEGAN-D2)的对比中可以看出.以 f-VAEGAN-D2 为例,平均提升效果 27.0%.由于直推式模型的参数更具泛化性,这类模型在 GZSL 任务下也取得了不错的效果,这从表 5 对应模型的效果可以看出.仍以 f-VAEGAN-D2 为例,平均提升效果 34.2%.

(3) 从打破潜在假设的角度来看,这类模型(表格中字体加粗的模型:RKT,CPL 和 BMVSc*)大致保持了已有工作实验效果,但能够使得模型的应用场景更加贴近实际.

在引入外部知识的模型中,表 4 中文献[119]的模型实验效果表明了:挖掘外部描述来替代人工标注语义的模型,经过噪声抑制等措施的处理,同样也能实现较好的效果.而引入外部知识库的模型,借助现有的知识图谱,其最大的优势在于可以方便地扩展到大规模的数据集中,实验效果见表 6.

Table 6 Experiments of different models on ImageNet21k, measuring top-1 accuracy (%)

表 6 各个模型在 ImageNet21k 数据集中的 Top-1 准确率(%)

模型	ZSL			GZSL		
	2H	3H	All	2H+1K	3H+1K	All+1K
EXEM[89,200]	12.5	3.6	1.8	4.3	1.3	0.7
GCNZ[130]	19.9	4.1	1.8	9.7	2.2	1.0
ADGPM[131]	26.6	6.3	3.0	10.3	2.9	1.4

表 6 中的 1K、2H、3H 和 ALL 分别表示训练集包含的 1 000 个类别、以训练集为核心的 2-hop 距离的类

别(借助相关 KG 的显式关系)、以训练集为核心的 3-hop 距离的类别、除训练集之外的所有类别.从中我们可以看到:GCNZ 和 ADGPM 这种引入了 KG 的模型,效果相较于之前的传统方法(EXEM 模型)均有较大的提升,效果平均提升 84.4%,充分说明了引入外部知识的有效性和必要性.

5 挑战与展望

- 预训练增益

随着深度网络架构的成熟,很多模型直接利用预训练好的 CNN 网络来进行目标数据集中样本视觉特征的提取.但如果预训练 CNN 的数据类别与目标数据集中不可见类部分有重叠,那么就会给零样本模型对不可见类数据的识别效果带来某些提升(增益),因此在 Xian 等人[4]的工作中,对现有的数据集进行了重新划分(PS 的划分方法),来避免这种情况的发生.也可以通过使用目标数据集中的可见类对预训练好的 CNN 进行微调来避免预训练增益情况的发生[112].但是,如何更好地防止增益效果对实验结果的影响,这是未来研究中需要注意的问题.

- 大数据集的挑战

ImageNet 21k 数据集,其庞大的数据规模中存在着较大的类别不均衡问题,是当前零样本图像分类任务中最具挑战性的数据集.文献[4]中对之前代表性的方法进行了集中的验证,近期工作也有少量模型在该数据集上进行了验证[89,103,112,130,131,200].在未来工作中,需要将这个数据集作为衡量零样本图像分类模型性能的基准.

- 伪样本的生成

利用生成式模型建模,在近两三年变得火热.生成更好的伪样本,有利于使用更少的数据来更快地确定准确的分类边界.但是如何生成质量更高的伪样本,这是生成式的方法面临的主要问题之一.基于 Wang 等人[8]工作中的表述,伪样本的生成应该有 3 个标准:真实性、有效性和多样性.其中,真实性是指生成的伪样本在视觉上要尽可能地接近真实的样本;有效性是指生成的目标类伪样本需要有利于目标类分类器的训练;多样性是指生成的某个类的伪样本,其类内方差要尽可能地大一些,更具有判别性.由于在 ZSL 领域,生成式模型是基于样本的高级特征来进行操作的,因此生成伪样本的真实性则更加适合于文本生成图像[187]等任务.而伪样本有效性方面需要优先进行考虑,通过在生成模型中引入条件,然后进行伪样本置信度筛选[83,104]等操作,伪样本的有效性有了较大的提升.在此基础上,增加生成伪样本的多样性,以更好地确定分类边界,是当前该类模型面临的挑战.文献[83,90]均通过类原型修正来提升所生成伪样本的多样性和区分性.同时,生成式模型生成伪样本的能力是有限的,如何将生成式模型进一步扩展到大规模数据集中(如 ImageNet21k),也是一个值得思考的问题.

- 模型的可解释性

深度网络的可解释性是近期比较热门的话题,但这里所指的可解释性是指从视觉角度出发,去阐述模型在进行零样本认知时的视觉依据,是一种弱可解释性.Xian 等人[112]的工作已对此进行了尝试.在未来的工作中,如何进一步扩展视觉可解释性的功能,甚至利用视觉层面的解释性来反馈辅助模型的训练,也是未来面临的挑战.

- 多义性

这一问题特指在词嵌入过程中发生的问题,即在词嵌入过程中,出现的一个词嵌入对应多个名词表示的现象.放在知识库中,即转化为“一个词嵌入对应多个图节点的表示”.在 WordNet 知识库中存在大量的这种现象(例如上下义词共享词嵌入表示),这是由于知识库的粒度和词嵌入表示的粒度不对等造成的.因此,这一问题更多是与知识表示是否合理有关.而多义性的存在是否会影响引入外部知识库的零样本模型性能,这也需要进行深入的探究.

6 结束语

在计算机视觉领域,由于数据爆发式增长带来信息标注成本高昂的问题,零样本学习越来越受到人们的重视.而随着“数据+知识驱动”这一理念深入到深度学习的各个领域,零样本学习也进入到新的发展阶段.本文针对当前对于“知识”这一概念并无统一表述的问题,对零样本学习领域所使用的知识进行了总结和归纳,从模型所使用的不同层次知识的角度出发,梳理了已有视觉相关的零样本学习工作(主要聚焦于零样本图像分类任

务);接着阐述了本研究领域的现存挑战,并基于存在挑战对已有工作进行了优缺点归纳;然后介绍了领域内常用数据、评估标准、实验分析;最后对未来工作进行了展望.本文的角度有助于人们理解零样本学习中的 3 大关键问题:如何更好地挖掘已知类的知识、如何更好地将获取的知识用于对未知类的认知中以及怎样合理地使用先验知识.

References:

[1] Zhou ZH. A brief introduction to weakly supervised learning. National Science Review, 2018,5(1):44−53.

[2] Thrun S, Pratt L. Learning to learn: Introduction and overview. In: Proc. of the Learning to Learn. Springer-Verlag, 1998. 3–17.

[3] Fu Y, 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.

[4] Xian Y, Lampert CH, Schiele B, *et al*. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,41(9):2251−2265.

[5] Wang W, Zheng VW, Yu H, *et al*. A survey of zero-shot learning: Settings, methods, and applications. ACM Trans. on Intelligent Systems and Technology (TIST), 2019,10(2):1–37.

[6] Wang Y, Yao Q, Kwok JT, *et al*. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys (CSUR), 2020,53(3):1–34.

[7] Larochelle H, Erhan D, Bengio Y. Zero-data learning of new tasks. In: Proc. of the AAAI. 2008. 646–651.

[8] Wang YX, Girshick R, Hebert M, *et al*. Low-shot learning from imaginary data. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7278−7286.

[9] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2009,22(10):1345−1359.

[10] Ye M, Guo Y. Progressive ensemble networks for zero-shot recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 11728−11736.

[11] Kodirov E, Xiang T, Fu Z, *et al*. Unsupervised domain adaptation for zero-shot learning. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 2452−2460.

[12] Zhao A, Ding M, Guan J, *et al*. Domain-invariant projection learning for zero-shot recognition. In: Advances in Neural Information Processing Systems. 2018. 1019−1030.

[13] Fu Y, Hospedales TM, Xiang T, *et al*. Transductive multi-view zero-shot learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,37(11):2332−2345.

[14] Fu Z, Xiang T, Kodirov E, *et al*. Zero-shot learning on semantic class prototype graph. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,40(8):2009−2022.

[15] Dinu G, Lazaridou A, Baroni M. Improving zero-shot learning by mitigating the hubness problem. arXiv preprint arXiv:1412.6568, 2014.

[16] Rohrbach M, Stark M, Schiele B. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: Proc. of the CVPR 2011. 2011. 1641−1648.

[17] Kumar N, Berg AC, Belhumeur PN, *et al*. Attribute and simile classifiers for face verification. In: Proc. of the 12th IEEE Int'l Conf. on Computer Vision. 2009. 365−372.

[18] Xiong F, Abdalmageed W. Unknown presentation attack detection with face RGB images. In: Proc. of the 9th IEEE Int'l Conf. on Biometrics Theory, Applications and Systems (BTAS). 2018. 1–9.

[19] Liu Y, Stehouwer J, Jourabloo A, *et al*. Deep tree learning for zero-shot face anti-spoofing. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 4680−4689.

[20] Liu J, Kuipers B, Savarese S. Recognizing human actions by attributes. In: Proc. of the CVPR 2011. 2011. 3337−3344.

[21] Cheng HT, Sun FT, Griss M, *et al*. Nuactiv: Recognizing unseen new activities using semantic attribute-based learning. In: Proc. of the 11th Annual Int'l Conf. on Mobile Systems, Applications, and Services. 2013. 361−374.

[22] Antol S, Zitnick CL, Parikh D. Zero-shot learning via visual abstraction. In: Proc. of the European Conf. on Computer Vision. 2014. 401−416.

[23] Jain M, Van Gemert JC, Mensink T, *et al*. Objects2action: Classifying and localizing actions without any video example. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4588−4596.

[24] Gan C, Lin M, Yang Y, *et al*. Exploring semantic inter-class relationships (SIR) for zero-shot action recognition. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2015. 3769−3775.

[25] Xu X, Hospedales TM, Gong S. Multi-task zero-shot action recognition with prioritised data augmentation. In: Proc. of the European Conf. on Computer Vision. 2016. 343−359.

[26] Gan C, Lin M, Yang Y, *et al*. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2016. 3487−3493.

[27] Wang W, Miao C, Hao S. Zero-shot human activity recognition via nonlinear compatibility based method. In: Proc. of the Int'l Conf. on Web Intelligence. 2017. 322−330.

[28] Xu X, Hospedales T, Gong SG. Transductive zero-shot action recognition by word-vector embedding. Int'l Journal of Computer Vision, 2017,123(3):309−333.

[29] Qin J, Liu L, Shao L, *et al*. Zero-shot action recognition with error-correcting output codes. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2833−2842.

[30] Mishra A, Verma VK, Reddy MSK, *et al*. A generative approach to zero-shot and few-shot action recognition. In: Proc. of the 2018 IEEE Winter Conf. on Applications of Computer Vision (WACV). 2018. 372−380.

[31] Gao J, Zhang T, Xu C. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2019. 8303−8311.

[32] Mandal D, Narayan S, Dwivedi SK, *et al*. Out-of-distribution detection for generalized zero-shot action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 9985−9993.

[33] Lampert CH, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,36(3):453−465.

[34] Deng J, Ding N, Jia Y, *et al*. Large-scale object classification using label relation graphs. In: Proc. of the European Conf. on Computer Vision. 2014. 48−64.

[35] Li LJ, Su H, Lim Y, *et al*. Objects as attributes for scene classification. In: Proc. of the European Conf. on Computer Vision. 2010. 57−69.

[36] Kordumova S, Mensink T, Snoek CG. Pooling objects for recognizing scenes without examples. In: Proc. of the 2016 ACM Int'l Conf. on Multimedia Retrieval. 2016. 143−150.

[37] Bucher M, Tuan-Hung VU, Cord M, *et al*. Zero-shot semantic segmentation. In: Advances in Neural Information Processing Systems. 2019. 468−479.

[38] Shen Y, Liu L, Shen F, *et al*. Zero-shot sketch-image hashing. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 3598−3607.

[39] Yelamarthi SK, Reddy SK, Mishra A, *et al*. A zero-shot framework for sketch based image retrieval. In: Proc. of the European Conf. on Computer Vision. 2018. 316−333.

[40] Dutta A, Akata Z. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 5089−5098.

[41] Fu Y, Hospedales TM, Xiang T, *et al*. Attribute learning for understanding unstructured social activity. In: Proc. of the European Conf. on Computer Vision. 2012. 530−543.

[42] Fu Y, Hospedales TM, Xiang T, *et al*. Learning multimodal latent attributes. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,36(2):303−316.

[43] Xu B, Fu Y, Jiang YG, *et al*. Video emotion recognition with transferred deep feature encodings. In: Proc. of the 2016 ACM Int'l Conf. on Multimedia Retrieval. 2016. 15−22.

[44] Zhang CR, Peng YX. Visual data synthesis via GAN for zero-shot video classification. arXiv preprint arXiv:1804.10073, 2018.

[45] Bansal A, Sikka K, Sharma G, *et al*. Zero-shot object detection. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 384−400.

[46] Lu C, Krishna R, Bernstein M, *et al*. Visual relationship detection with language priors. In: Proc. of the European Conf. on Computer Vision. 2016. 852−869.

[47] Dalton J, Allan J, Mirajkar P. Zero-shot video retrieval using content and concepts. In: Proc. of the 22nd ACM Int'l Conf. on Information and Knowledge Management. 2013. 1857−1860.

[48] Wu S, Bondugula S, Luisier F, *et al*. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 2665−2672.

[49] Chang X, Yang Y, Hauptmann A, *et al*. Semantic concept discovery for large-scale zero-shot event detection. In: Proc. of the 24th Int'l Joint Conf. on Artificial Intelligence. 2015. 2234−2240.

[50] Chang X, Yang Y, Long G, *et al*. Dynamic concept composition for zero-example event detection. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2016. 3464−3470.

[51] Blitzer J, Foster DP, Kakade SM. Zero-shot domain adaptation: A multi-view approach. Technical Report, TTI-TR-2009-1, Toyota Technological Institute at Chicago, 2009.

[52] Yazdani M, Henderson J. A model of zero-shot learning of spoken language understanding. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015. 244−249.

[53] Jiang H, Wang R, Shan S, *et al*. Learning class prototypes via structure alignment for zero-shot recognition. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 118−134.

[54] Parikh D, Grauman K. Relative attributes. In: Proc. of the 2011 Int'l Conf. on Computer Vision. 2011. 503−510.

[55] Parikh D, Grauman K. Interactively building a discriminative vocabulary of nameable attributes. In: Proc. of the CVPR 2011. 2011. 1681−1688.

[56] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[57] Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. 2013. 3111−3119.

[58] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1532−1543.

[59] Fu Y, Hospedales TM, Xiang T, *et al*. Transductive multi-view embedding for zero-shot recognition and annotation. In: Proc. of the European Conf. on Computer Vision. 2014. 584−599.

[60] Fu Z, Xiang T, Kodirov E, *et al*. Zero-shot object recognition by semantic manifold distance. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2635−2644.

[61] Fu Y, Sigal L. Semi-supervised vocabulary-informed learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5337−5346.

[62] Zhang C, Lyu X, Tang Z. TGG: Transferable graph generation for zero-shot and few-shot learning. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. 2019. 1641−1649.

[63] Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 951−958.

[64] Huang S, Elhoseiny M, Elgammal A, *et al*. Learning hypergraph-regularized attribute predictors. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 409−417.

[65] Ye M, Guo Y. Zero-shot classification with discriminative semantic representation learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 7140−7148.

[66] Kodirov E, Xiang T, Gong S. Semantic autoencoder for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3174−3183.

[67] Lu Z, Guan J, Li A, *et al*. Zero and few shot learning with semantic feature synthesis and competitive learning. arXiv preprint arXiv:1810.08332, 2018.

[68] Zhang L, Xiang T, Gong S. Learning a deep embedding model for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2021−2030.

[69] Liu S, Long M, Wang J, *et al*. Generalized zero-shot learning with deep calibration network. In: Advances in Neural Information Processing Systems. 2018. 2005−2015.

[70] Liu ZZ, Zhang XX, Zhu ZF, *et al*. Convolutional prototype learning for zero-shot recognition. arXiv preprint arXiv:1910.09728, 2019.

[71] Jiang H, Wang R, Shan S, *et al*. Learning discriminative latent attributes for zero-shot classification. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 4223−4232.

[72] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012. 1097−1105.

[73] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[74] van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008,9:2579−2605.

[75] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. 2017. 4077−4087.

[76] Zhou L, Cui P, Yang S, *et al*. Learning to learn image classifiers with informative visual analogy. arXiv preprint arXiv:1710.06177, 2017.

[77] Wan Z, Chen D, Li Y, *et al*. Transductive zero-shot learning with visual structure constraint. In: Advances in Neural Information Processing Systems. 2019. 9972−9982.

[78] Zhu Y, Xie J, Tang Z, *et al*. Semantic-guided multi-attention localization for zero-shot learning. In: Advances in Neural Information Processing Systems. 2019. 14917−14927.

[79] Li Y, Wang D, Hu H, *et al*. Zero-shot recognition using dual visual-semantic mapping paths. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3279−3287.

[80] Wang X, Pang S, Zhu J, *et al*. Visual space optimization for zero-shot learning. arXiv preprint arXiv:1907.00330, 2019.

[81] Socher R, Ganjoo M, Manning CD, *et al*. Zero-shot learning through cross-modal transfer. In: Advances in Neural Information Processing Systems. 2013. 935−943.

[82] Li Y, Zhang J, Zhang J, *et al*. Discriminative learning of latent features for zero-shot recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7463−7471.

[83] Li J, Jin M, Lu K, *et al*. Leveraging the invariant side of generative zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 7402−7411.

[84] Zhu P, Wang H, Saligrama V. Generalized zero-shot recognition based on visually semantic embedding. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 2995−3003.

[85] Zhao B, Wu B, Wu T, *et al*. Zero-shot learning posed as a missing data problem. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2616−2622.

[86] Kumar Verma V, Arora G, Mishra A, *et al*. Generalized zero-shot learning via synthesized examples. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 4281−4289.

[87] Changpinyo S, Chao WL, Gong B, *et al*. Classifier and exemplar synthesis for zero-shot learning. Int'l Journal of Computer Vision, 2020,128(1):166−201.

[88] Long Y, Liu L, Shao L, *et al*. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1627−1636.

[89] Changpinyo S, Chao WL, Sha F. Predicting visual exemplars of unseen classes for zero-shot learning. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 3476−3485.

[90] Liu B, Dong QL, Hu ZY. Zero-shot learning from adversarial feature residual to compact visual feature. In: Proc. of the AAAI. 2020. 11547−11554.

[91] Zhang Z, Saligrama V. Zero-shot recognition via structured prediction. In: Proc. of the European Conf. on Computer Vision. 2016. 533−548.

[92] Xu X, Shen F, Yang Y, *et al*. Matrix tri-factorization with manifold regularizations for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3798−3807.

[93] Xu X, Shen F, Yang Y, *et al*. Transductive visual-semantic embedding for zero-shot learning. In: Proc. of the 2017 ACM Int'l Conf. on Multimedia Retrieval. 2017. 41−49.

[94] Wang Q, Chen K. Zero-shot visual recognition via bidirectional latent embedding. Int'l Journal of Computer Vision, 2017,124(3): 356−383.

[95] Li Y, Jia Z, Zhang J, *et al*. Deep semantic structural constraints for zero-shot learning. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018. 7049−7056.

[96] Changpinyo S, Chao WL, Gong B, *et al*. Synthesized classifiers for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5327−5336.

[97] Mukherjee T, Hospedales T. Gaussian visual-linguistic embedding for zero-shot recognition. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. 2016. 912−918.

[98] Guo Y, Ding G, Han J, *et al*. Synthesizing samples fro zero-shot learning. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence. 2017. 1774−1780.

[99] Bucher M, Herbin S, Jurie F. Generating visual representations for zero-shot classification. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 2666−2673.

[100] Verma VK, Rai P. A simple exponential family framework for zero-shot learning. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. 2017. 792−808.

[101] Wang W, Pu Y, Verma VK, *et al*. Zero-shot learning via class-conditioned deep generative models. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2018. 4211−4218.

[102] Mishra A, Reddy SK, Mittal A, *et al*. A generative model for zero shot learning using conditional variational autoencoders. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops. 2018. 2188−2196.

[103] Schonfeld E, Ebrahimi S, Sinha S, *et al*. Generalized zero-and few-shot learning via aligned variational autoencoders. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 8247−8255.

[104] Yu H, Lee B. Zero-shot learning via simultaneous generating and learning. In: Advances in Neural Information Processing Systems. 2019. 46−56.

[105] Tong B, Klinkigt M, Chen J, *et al*. Adversarial zero-shot learning with semantic augmentation. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. 2018.

[106] Xian Y, Lorenz T, Schiele B, *et al*. Feature generating networks for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 5542−5551.

[107] Zhu Y, Elhoseiny M, Liu B, *et al*. A generative adversarial approach for zero-shot learning from noisy texts. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1004−1013.

[108] Paul A, Krishnan NC, Munjal P. Semantically aligned bias reducing zero shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 7056−7065.

[109] Ni J, Zhang S, Xie H. Dual adversarial semantics-consistent network for generalized zero-shot learning. In: Advances in Neural Information Processing Systems. 2019. 6143−6154.

[110] Felix R, Kumar VB, Reid I, *et al*. Multi-modal cycle-consistent generalized zero-shot learning. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 21−37.

[111] Huang H, Wang C, Yu PS, *et al*. Generative dual adversarial network for generalized zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 801−810.

[112] Xian Y, Sharma S, Schiele B, *et al*. f-VAEGAN-D2: A feature generating framework for any-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 10275−10284.

[113] Liu H, Zheng QH, Luo MN, Zhao HK, Xiao Y, Lü YZ. Cross-domain adversarial learning for zero-shot classification. Journal of Computer Research and Development, 2019,56(12):2521−2535 (in Chinese with English abstract).

[114] Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013.

[115] Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. In: Advances in Neural Information Processing Systems. 2014. 2672−2680.

[116] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[117] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv preprint arXiv:1701.07875, 2017.

[118] Ba JL, Swersky K, Fidler S, *et al*. Predicting deep zero-shot convolutional neural networks using textual descriptions. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4247−4255.

[119] Qiao R, Liu L, Shen C, *et al*. Less is more: Zero-shot learning from online textual documents with noise suppression. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2249−2257.

[120] Elhoseiny M, Zhu Y, Zhang H, *et al*. Link the head to the "beak": Zero shot learning from noisy text description at part precision. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017. 6288−6297.

[121] Speer R, Havasi C. ConceptNet 5: A large semantic network for relational knowledge. In: Proc. of the People's Web Meets NLP. Springer-Verlag, 2013. 161−176.

[122] Speer R, Chin J, Havasi C. Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. 2017. 4444−4451.

[123] Liu ZY, Sun MS, Lin YK, Xie RB. Knowledge representation learning: A review. Journal of Computer Research and Development, 2016,53(2):247−261 (in Chinese with English abstract).

[124] Scarselli F, Gori M, Tsoi AC, *et al*. The graph neural network model. IEEE Trans. on Neural Networks, 2008,20(1):61−80.

[125] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[126] Al-Halah Z, Stiefelhagen R. How to transfer? Zero-shot object recognition via hierarchical transfer of semantic attributes. In: Proc. of the 2015 IEEE Winter Conf. on Applications of Computer Vision. 2015. 837−843.

[127] Li X, Liao S, Lan W, *et al*. Zero-shot image tagging by hierarchical semantic embedding. In: Proc. of the 38th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 2015. 879−882.

[128] Li AX, Zhang KX, Wang LW. Zero-shot fine-grained classification by deep feature learning with semantics. Int'l Journal of Automation and Computing, 2019,16(5):563−574.

[129] Cui P, Liu SW, Zhu WW. General knowledge embedded image representation learning. IEEE Trans. on Multimedia, 2017,20(1): 198−207.

[130] Wang X, Ye Y, Gupta A. Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6857−6866.

[131] Kampffmeyer M, Chen Y, Liang X, *et al*. Rethinking knowledge graph propagation for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 11487−11496.

[132] Lee CW, Fang W, Yeh CK, *et al*. Multi-label zero-shot learning with structured knowledge graphs. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1576−1585.

[133] Deselaers T, Ferrari V. Visual and semantic similarity in imagenet. In: Proc. of the CVPR 2011. 2011. 1777−1784.

[134] Farhadi A, Endres I, Hoiem D, *et al*. Describing objects by their attributes. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 1778−1785.

[135] Cheng YH, Qiao X, Wang XS. Hybrid attribute-based zero-shot image classification. Acta Electronica Sinica, 2017,45(6): 1462−1468 (in Chinese with English abstract).

[136] Turakhia N, Parikh D. Attribute dominance: What pops out? In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 1225−1232.

[137] Suzuki M, Sato H, Oyama S, *et al*. Transfer learning based on the observation probability of each attribute. In: Proc. of the 2014 IEEE Int'l Conf. on Systems, Man, and Cybernetics (SMC). 2014. 3627−3631.

[138] Jayaraman D, Grauman K. Zero-shot recognition with unreliable attributes. In: Advances in Neural Information Processing Systems. 2014. 3464−3472.

[139] Rohrbach M, Stark M, Szarvas G, *et al*. What helps where—And why? Semantic relatedness for knowledge transfer. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2010. 910−917.

[140] Yu X, Aloimonos Y. Attribute-based transfer learning for object categorization with zero/one training example. In: Proc. of the European Conf. on Computer Vision. 2010. 127−140.

[141] Wang X, Ji Q. A unified probabilistic approach modeling relationships between attributes and objects. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 2120−2127.

[142] Hariharan B, Vishwanathan SVN, Varma M. Efficient max-margin multi-label classification with applications to zero-shot learning. Machine learning, 2012,88(1-2):127−155.

[143] Kankuekul P, Kawewong A, Tangruamsub S, *et al*. Online incremental attribute-based zero-shot learning. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. 2012. 3657−3664.

[144] Morgado P, Vasconcelos N. Semantically consistent regularization for zero-shot recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6060−6069.

[145] Lu J, Li J, Yan Z, *et al*. Attribute-based synthetic network (ABS-net): Learning more from pseudo feature representations. Pattern Recognition, 2018,80:129−142.

[146] Jayaraman D, Sha F, Grauman K. Decorrelating semantic visual attributes by resisting the urge to share. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 1629−1636.

[147] Liang K, Chang H, Shan S, *et al*. A unified multiplicative framework for attribute learning. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 2506−2514.

[148] Gan C, Yang T, Gong B. Learning attributes equals multi-source domain generalization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 87−97.

[149] Li H, Li D, Luo X. Bap: Bimodal attribute prediction for zero-shot image categorization. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. 2014. 1013−1016.

[150] Yu FX, Cao L, Feris RS, *et al*. Designing category-level attributes for discriminative visual recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 771−778.

[151] Wang D, Li Y, Lin Y, *et al*. Relational knowledge transfer for zero-shot learning. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. 2016.

[152] Lazaridou A, Bruni E, Baroni M. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol.1: Long Papers). 2014. 1403−1414.

[153] Guo Y, Ding G, Jin X, *et al*. Transductive zero-shot recognition via shared model space learning. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. 2016.

[154] Romera-Paredes B, Torr P. An embarrassingly simple approach to zero-shot learning. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 2152−2161.

[155] Lazaridou A, Dinu G, Baroni M. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing (Vol.1: Long Papers). 2015. 270−280.

[156] Atzmon Y, Chechik G. Adaptive confidence smoothing for generalized zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 11671−11680.

[157] Pasupat P, Liang P. Zero-shot entity extraction from Web pages. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol.1: Long Papers). 2014. 391−401.

[158] Shigeto Y, Suzuki I, Hara K, *et al*. Ridge regression, hubness, and zero-shot learning. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. 2015. 135−151.

[159] Frome A, Corrado GS, Shlens J, *et al*. Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems. 2013. 2121−2129.

[160] Annadani Y, Biswas S. Preserving semantic relations for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7603−7612.

[161] Chen L, Zhang H, Xiao J, *et al*. Zero-Shot visual recognition using semantics-preserving adversarial embedding networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1043−1052.

[162] Tong B, Wang C, Klinkigt M, *et al*. Hierarchical disentanglement of discriminative latent features for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 11467−11476.

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

[164] Zhang Z, Saligrama V. Zero-shot learning via semantic similarity embedding. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4166−4174.

[165] Demirel B, Gokberk Cinbis R, Ikizler-Cinbis N. Attributes2Classname: A discriminative model for attribute-based unsupervised zero-shot learning. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1232−1241.

[166] Reed S, Akata Z, Lee H, *et al*. Learning deep representations of fine-grained visual descriptions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 49−58.

[167] Ding Z, Shao M, Fu Y. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2050−2058.

[168] Ji Z, Fu Y, Guo J, *et al*. Stacked semantics-guided attention model for fine-grained zero-shot learning. In: Advances in Neural Information Processing Systems. 2018. 5995−6004.

[169] Akata Z, Perronnin F, Harchaoui Z, *et al*. Label-embedding for attribute-based classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 819−826.

[170] Akata Z, Reed S, Walter D, *et al*. Evaluation of output embeddings for fine-grained image classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2927−2936.

[171] Xian Y, Akata Z, Sharma G, *et al*. Latent embeddings for zero-shot classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 69−77.

[172] Zhang Z, Saligrama V. Zero-shot learning via joint latent similarity embedding. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 6034−6042.

[173] Yu Y, Ji Z, Guo J, *et al*. Transductive zero-shot learning with adaptive structural embedding. IEEE Trans. on Neural Networks and Learning Systems, 2017,29(9):4116−4127.

[174] Jiang H, Wang R, Shan S, *et al*. Adaptive metric learning for zero-shot recognition. IEEE Signal Processing Letters, 2019,26(9): 1270−1274.

[175] Song J, Shen C, Yang Y, *et al*. Transductive unbiased embedding for zero-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1024−1033.

[176] Zheng J, Zhuang F, Shi C. Local ensemble across multiple sources for collaborative filtering. In: Proc. of the 2017 ACM on Conf. on Information and Knowledge Management. 2017. 2431−2434.

[177] Tsai YHH, Huang LK, Salakhutdinov R. Learning robust visual-semantic embeddings. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). 2017. 3591−3600.

[178] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[179] Jiang H, Wang R, Shan S, *et al*. Transferable contrastive network for generalized zero-shot learning. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 9765−9774.

[180] Sung F, Yang Y, Zhang L, *et al*. Learning to compare: Relation network for few-shot learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1199−1208.

[181] Hu RL, Xiong C, Socher R. Correction networks: Meta-learning for zero-shot learning. In: Proc. of the ICLR. 2019. 1−12.

[182] Akata Z, Perronnin F, Harchaoui Z, *et al*. Label-embedding for image classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015,38(7):1425−1438.

[183] Misra I, Gupta A, Hebert M. From red wine to red tomato: Composition with context. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1792−1801.

[184] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006,7(85):2399−2434.

[185] Vilnis L, McCallum A. Word representations via Gaussian embedding. arXiv preprint arXiv:1412.6623, 2014.

[186] Micaelli P, Storkey AJ. Zero-shot knowledge transfer via adversarial belief matching. In: Advances in Neural Information Processing Systems. 2019. 9551−9561.

[187] Reed S, Akata Z, Yan X, *et al*. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396, 2016.

[188] Hendricks LA, Akata Z, Rohrbach M, *et al*. Generating visual explanations. In: Proc. of the European Conf. on Computer Vision. 2016. 3−19.

[189] Mitchell T, Cohen W, Hruschka E, *et al*. Never-ending learning. Communications of the ACM, 2018,61(5):103−115.

[190] Chen X, Shrivastava A, Gupta A. Neil: Extracting visual knowledge from Web data. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 1409−1416.

[191] Li Y, Tarlow D, Brockschmidt M, *et al*. Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493, 2015.

[192] Marino K, Salakhutdinov R, Gupta A. The more you know: Using knowledge graphs for image classification. arXiv preprint arXiv:1612.04844, 2016.

[193] Low T, Borgelt C, Stober S, *et al*. The hubness phenomenon: Fact or artifact? In: Towards Advanced Data Analysis by Combining Soft Computing and Statistics. Berlin, Heidelberg: Springer-Verlag, 2013. 267−278.

[194] Wah C, Branson S, Welinder P, *et al*. The Caltech-UCSD Birds-200-2011 dataset. Technical Report, 2011.

[195] Patterson G, Hays J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. 2012. 2751−2758.

[196] Nilsback ME, Zisserman A. Automated flower classification over a large number of classes. In: Proc. of the 2008 6th Indian Conf. on Computer Vision, Graphics and Image Processing. 2008. 722−729.

[197] Xiao J, Hays J, Ehinger KA, *et al*. Sun database: Large-scale scene recognition from abbey to zoo. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2010. 3485−3492.

[198] Elhoseiny M, Saleh B, Elgammal A. Write a classifier: Zero-shot learning using purely textual descriptions. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 2584−2591.

[199] Deng J, Dong W, Socher R, *et al*. Imagenet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 248−255.

[200] Chao WL, Changpinyo S, Gong B, *et al*. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: Proc. of the European Conf. on Computer Vision. 2016. 52−68.

附中文参考文献:

[113] 刘欢,郑庆华,罗敏楠,赵洪科,肖阳,吕彦章.基于跨域对抗学习的零样本分类.计算机研究与发展,2019,56(12):2521−2535.

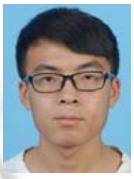
[123] 刘知远,孙茂松,林衍凯,谢若冰.知识表示学习研究进展.计算机研究与发展,2016,53(2):247−261.

[135] 程玉虎,乔雪,王雪松.基于混合属性的零样本图像分类.电子学报,2017,45(6):1462−1468.

冯耀功(1992－),男,博士生,CCF 学生会员,主要研究领域为深度学习,计算机视觉,零样本学习.

桑基韬(1985－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为多媒体计算,网络数据挖掘,可信赖机器学习.

于剑(1969－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为人工智能,机器学习.



杨朋波(1993－),男,博士生,CCF 学生会员,主要研究领域为深度学习,计算机视觉,对抗鲁棒性.