零样本图像分类综述: 十年进展

冀中, 汪浩然, 于云龙, 庞彦伟*

天津大学电气自动化与信息工程学院, 天津 300072

* 通信作者. E-mail: pyw@tju.edu.cn

收稿日期: 2019-03-05; 修回日期: 2019-05-25; 接受日期: 2019-06-03; 网络出版日期: 2019-10-16

国家自然科学基金 (批准号: 61171329, 61632018) 资助项目

摘要 零样本图像分类指训练集和测试集在数据的类别上没有交集的情况下进行图像分类. 该技术是为解决类别标签缺失问题的一种有效手段, 因此受到了日益广泛的关注. 自提出此问题至今, 零样本图像分类的研究已经大致有十年时间. 本文系统地对过去十年中零样本图像分类技术的研究进展进行了综述, 主要包括以下 4 个方面. 首先介绍零样本图像分类技术的研究意义及其应用价值, 然后重点总结和归纳零样本图像分类的发展过程和研究现状, 接下来介绍常用的数据集和评价准则, 以及与零样本学习相关的技术的区别和联系, 最后分析有待深入研究的热点与难点问题, 并对未来的发展趋势进行了展望.

关键词 零样本图像分类, 属性, 词向量, 跨模态映射, 领域适应学习

1 意义

零样本分类 (zero-shot classification, ZSC), 也称作零样本学习 (zero-shot learning), 是指依据一些可见类别 (seen classes) 的数据, 辅以相关常识信息或先验知识 (称为辅助信息, side information), 用于训练某种学习模型, 实现对未见类别 (unseen classes, 指数据标注完全缺失的类别) 的数据进行类别预测和识别的一类技术^[1~5]. 在这种分类场景下, 训练集合和测试集合的类别是不相交的, 这显著有别于传统的分类任务^[6], 如图 1 所示. 其中 y 和 z 分别表示训练样本和测试样本, $\mathcal{Y}$ 和 $\mathcal{Z}$ 分别表示训练数据和测试数据所属的类别. 零样本分类中测试样本所属的类别和训练数据所属的类别是不相交的, 而传统的目标分类中测试样本所属的类别包含于训练类别的集合.

现有的物体分类技术大多属于监督学习, 即需要大量标注数据训练相关模型^[7,8]. 据统计^[9], 人类可识别的物体种类大约有 3 万类, 要对如此之多的类别进行数据标注是一件极其费力的工作. 此外, 现实世界中物体的种类还在不断增长, 这需要识别系统不断增加新的数据并重新进行训练, 这些问题

引用格式: 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: 1299–1320, doi: 10.1360/N112018-00312
Ji Z, Wang H R, Yu Y L, et al. A decadal survey of zero-shot image classification (in Chinese). Sci Sin Inform, 2019, 49: 1299–1320, doi: 10.1360/N112018-00312

图 1 (网络版彩图) 零样本分类与传统的目标分类的区别

Figure 1 (Color online) Difference between zero-shot classification and traditional classification task. (a) Zero-shot classification; (b) traditional object classification.

严重阻碍了相关技术的进一步发展. 人们迫切需要一种在目标类别的视觉标注数据完全缺失的情况下仍然能够识别这些类别数据的技术. 受到现实需求以及技术不断发展的双重驱动, 零样本分类技术应运而生.

与传统的分类技术不同, 零样本分类技术很大程度上受到了人类学习过程的启发. 在学习机理上与传统的分类技术的不同, 决定了零样本分类必须走一条继承和发展的研究道路. 在 2016 年中国计算机大会上, 谭铁牛院士指出, 生物启发的模式识别是一个非常值得关注的研究方向, “比如人识别一个动物 (并不需要看到过该动物), 只需要一句话的描述就能识别出来该动物”. 例如, 假设一个小孩从未见过斑马 (斑马对小孩来说是未见类), 但他知道斑马是一种身上有黑白条纹的外形像马的动物 (马对于小孩来说是可见类), 那么当他第一次看到斑马的时候, 一般都能认出这是斑马. 这样的类似“触类旁通”学习过程可以总结为“已有学习模型 (认识马) + 常识或先验 (知道斑马是身上有黑白条纹的外形像马的动物) + 进行演绎推断 (识别斑马)”的过程, 这正是零样本分类的思想核心, 即利用常识或先验知识进行演绎推理.

零样本分类是目标分类技术发展的必然趋势之一. 一方面, 传统需要大量标注样本进行学习模式的不足迫使新的需要较少甚至无需标注样本技术的发展. 另一方面, 计算机视觉和机器学习等技术的发展在技术层面促进了零样本分类的产生和发展. 为此, 学者们在 2003 年前后提出了单样本学习 (one-shot learning/few-shot learning) 概念^[10,11], 目的是仅利用一个或者少量的标注样本对学习模型进行训练. 进一步, 在 2008 年前后又提出了零样本学习技术^[1~3], 目的是研究某些类别标注完全缺失的情况. 零样本分类可看作视觉数据与文本等其他模态数据间的一种跨模态学习^[12~14]. 这与人类的学习过程类似: 我们既不是从海量的图像或者视频数据集中学习相关类别概念, 也不是从海量语料库中学习语言 (语义), 而是从视觉和语言 (语义) 的交互中理解了世间万物. 零样本分类的技术目前正处于高速发展时期, 所涉及的具体应用已经从最初的图像分类任务扩展到了其他计算机视觉任务^[15~23]乃至自然语言处理^[24,25]等多个相关领域. 对此, 本文将其称为广义零样本分类. 相应地, 我们将针对图像分类任务的零样本分类任务称为狭义零样本分类. 需要指出的是, 由于本文聚焦于讨论狭义零样本分类技术, 如无特殊说明, 后文中出现的零样本分类均表示狭义零样本分类, 即零样本图像分类.

从理论体系从属关系的角度来看, 零样本分类属于迁移学习^[26~28]领域的一个子任务. 迁移学习的主要目的是将源域 (source domain) 和源任务 (source task) 中的知识迁移到目标域 (target domain) 中用于帮助目标任务 (target task) 进行学习. 对于零样本分类任务, 可见类和未见类的数据可以分别视为来自源域和目标域, 因此零样本分类可以看作是迁移学习的一种特殊应用. 另一方面, 从模型实现的角度来看, 零样本分类的实现与另外两个研究领域密不可分, 其一是表征学习 (representation

learning), 其二是度量学习 (metric learning). 要建立一个好的机器学习模型首先离不开优秀的数据表征能力. 表征学习是指通过对数据进行变换从而提取数据中的有效信息的一种学习方式, 涉及到人工智能相关的诸多领域, 如语音识别^[29]、信号处理^[30]、目标识别^[7]、自然语言处理^[31], 以及迁移学习^[32]等. 就零样本分类任务而言, 目前最主流的学习方式是利用嵌入模型 (详情见本文 2.2 小节), 即将多个模型的数据嵌入到同一语义空间中进行数据表征, 其本质可以视为一种以语义表示为核心的表征学习过程. 度量学习通常建立在表征学习的基础之上, 其本质是根据不同的任务, 根据特定空间中的数据自主学习出针对某个特定任务的距离度量函数, 目前已被广泛应用于诸多计算机视觉相关的任务, 如人脸识别^[33]、图像检索^[34]、目标跟踪^[35]、多模态匹配^[36]等. 对于零样本分类任务, 在获取到合适的数据表征空间之后, 则需要对跨模态样本间的距离度量进行学习, 目的是保证嵌入到语义空间后样本间的语义相似度关系得以保持. 因此, 零样本分类又可以看作是在不同域中的度量学习过程. 综上所述, 零样本学习可以看作是在进行表征学习和度量学习的基础上, 通过借助辅助信息 (属性或文本) 实现跨模态知识的迁移, 从而完成可见类信息到未见类信息推断的迁移学习过程.

相较于一些经典课题, 零样本分类技术的发展还不是很成熟. 目前, 关于零样本分类的综述文献较少, 其中影响力较大的分别是 Xian 等^[37]、Fu 等^[38], 以及 Wang 等^[39]的工作. 其中, Xian 等^[37]对目前零样本分类的部分经典分类模型做出了一些概括性总结, 并聚焦于目前诸多零样本分类工作评判标准各异的问题, 提出了一些统一的标准 (如性能评价指标, 数据集划分方式等), 并基于上述标准对一些分类模型进行了统一的性能测试, 供其他研究者参考. 文献^[38]则着重于对零样本分类及其相关领域进行更加全面的介绍, 除了分类模型、数据集、评价指标之外, 还对广义零样本分类任务进行了更加详细的介绍和讨论. 相较于前两者, 文献^[39]的主要贡献在于首次对零样本分类问题中不同学习方式的设定进行了正式的定义. 整体而言, 上述文献的共性是侧重于讨论零样本分类的发展现状. 而本文以时间为线索, 分析了零样本技术的前世今生、发展脉络, 以及技术要点, 讨论分析了与相关技术的区别联系, 并且对零样本分类技术面临的问题与挑战以及几个重要的发展趋势进行了探讨, 旨在让广大读者了解零样本技术的发展历史、常见技术, 以及发展趋势.

2 发展及现状分析

零样本分类是目标分类研究方向的一个分支, 其提出经历了从传统的目标分类到单样本学习, 再到零样本分类这几个阶段. 如图 2 所示. 图中不同的形状代表不同的类别, $\mathcal{Y}$ 和 $\mathcal{Z}$ 分别表示训练类别集合和测试类别集合. 传统的目标分类 (每一个训练类别有大量的标注样本) 和单样本学习 (每一个训练类别只有一个或少量的标注样本) 中测试类别集合属于训练类别集合, 即测试样本一定属于训练阶段出现过的类别; 零样本分类中测试类别集合和训练类别集合是不相交的, 即测试样本一定不属于训练过的类别; 而开集零样本分类中, 训练类别的集合包含于测试类别的集合, 即测试样本既可能属于已训练过的类别也可能属于未训练过的类别. 零样本分类处理的对象包括图像、文本和音频等, 本论文只介绍零样本图像分类技术 (视频视作一种特殊的图像).

对于现阶段的零样本分类任务, 成功的关键在于对视觉和语义这两种模态进行跨模态映射关系的学习. 从视觉特征和类别语义表征技术演进的时间角度来看, 本文将零样本分类技术的发展大致划分为两个阶段. 第 1 阶段可以称为提出阶段, 时间区间大致为 2008~2012 年, 这个阶段的主流研究思路是利用“浅层视觉特征 + 属性 + 传统机器学习方法”的分析模式. 第 2 阶段可以称为发展阶段, 时间区间大致为 2013~2019 年, 这个阶段的主流研究思路是利用“深度视觉特征 + 属性/词向量 + 传统机器学习方法”的分析模式. 这两个阶段总计经历了大约十年的时间, 在此期间出现的一些有代表性

图 2 (网络版彩图) 目标分类技术的发展过程和趋势
Figure 2 (Color online) The development and trend of object classification

的方法正是本文所要综述的重点内容. 本节将在 2.1 和 2.2 小节中以时间为线索分别对零样本分类技术的提出阶段和发展阶段的主要进展进行概述.

2.1 零样本分类技术的提出阶段

零样本分类技术的提出阶段大致在 2008~2012 年间, 主流研究思路是利用“低层视觉特征 + 属性 + 传统机器学习方法”的分析模式. 零样本分类的提出在很大程度上受到了人类学习过程的启发. 人类能够根据类别的描述对未见类别的样本进行推断和预测, 受此启发, 零样本分类通过可见类数据建立视觉特征和类别描述之间的语义关系, 借助常识或先验将知识从可见类别转移到未见类别, 以实现未见类别的样本的推断和预测. 其概念在 2008 年由 Larochelle 等^[1]首次提出, 其主要贡献在于将零样本分类公式化, 并利用实验验证了零样本分类在字符识别和多任务排序问题中的可行性和有效性.

随后, 在 2009 年涌现出 3 个有代表性的工作^[2,3,40]. 其中, Palatucci 等^[2]的工作, 首次提出了利用类别的语义描述对未知类别进行推断, 并在 PAC 框架下探索了能够对未见类别进行正确预测的条件. Farhadi 等^[40]首次将识别问题转变为属性描述问题, 并针对如何预测和分类未知类别的样本的问题进行了研究和探讨. 其中属性 (attribute) 作为一种先验知识能够对特定类别的物体进行类别语义表征, 它具有如下几个主要特点: (1) 是类别的一些特性; (2) 可命名 (如颜色、部位、习性等); (3) 具有判别性; (4) 可以通过视觉样本进行学习. 属性特征可以通过人为定义^[3]或通过属性学习^[41]的方法等多种途径获得.

在这一阶段, 属性作为对类别进行类别语义表征的主要方式, 为视觉和语义, 以及语义之间的关系建立了桥梁, 也是零样本分类技术得以提出的一个重要推动因素^[3,40,42~44].

2.2 零样本分类技术的发展阶段

零样本分类技术的发展大致时间为 2013 年至今, 主流研究思路是利用“深度视觉特征 + 属

性/词向量 + 传统机器学习方法”的分析模式. 在这一阶段, 受益于深度学习技术的发展, 特别是 2012 年卷积神经网络 (convolutional neural network, CNN) 模型 AlexNet^[7] 和 2013 年神经语言模型 Word2Vec^[31] 的提出, 零样本分类技术得到了较快的发展.

在视觉特征提取方面, 零样本分类中使用的视觉特征大多为使用 CNN 提取的深度视觉特征^[45~47] 或其改进^[48].

在类别的类别语义表征方面, 除属性学习的方式之外, 越来越多的学者尝试在无监督信息的条件下进行类别语义表征的学习. 这主要是因为通过属性学习进行类别语义表征的方式有着诸多局限性: (1) 属性学习需要对类别进行属性标注作为监督信息, 但属性标注需要较强的专业知识和较高的成本, 属性标注的质量对模型分类性能也会产生较大影响; (2) 属性标注一般是针对特定数据集的, 这意味着想要通过属性学习实现跨数据集的零样本分类 (特别是数据集间差异较大时) 几乎不可能. 因此在这一阶段, 无监督信息的类别语义表征方法成为了新的研究热点. 其中, 最重要的当属 Mikolov 等^[31] 提出的词向量 (Word2Vec) 技术. 作为一种新的包含常识性语义关系信息的语义表征技术, 词向量自从提出至今始终备受学界的关注并得到广泛使用. 词向量是指利用自然语言处理技术将文本表示为语义向量的一类技术的总称. 常用方法包括 CBOW^[31]、Skip-gram^[31]、GloVe^[49] 等基于神经语言模型的方法, 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.

除了词向量技术, 也有学者开始尝试其他无监督的类别语义表征方法, 如 Reed 等^[50] 使用文本描述 (text description) 进行类别语义表征, Karessli 等^[51] 则首次使用人眼的关注区域 (human gaze) 进行类别语义表征. 值得关注的是, 受益于深度学习技术的快速发展, 零样本分类技术中的视觉特征提取能力和类别语义表征能力均得到了显著提高, 但目前大多数性能较好的零样本分类算法的分类模型仍使用传统的机器学习方法.

2.3 小结

本节以时间顺序为线索, 将现有的零样本分类方法划分为提出和发展两个阶段, 并对各自发展阶段中出现的一些重要的工作进行了简要介绍和总结, 旨在帮助读者快速了解零样本分类技术的发展脉络. 第 3 节仍将以时间为主线, 进一步对零样本分类的技术体系进行系统梳理并对其中的诸多代表性工作进行介绍.

3 零样本分类技术的体系梳理

第 2 节主要以发展阶段 (即时间顺序) 为主线对零样本分类技术进行了简要总结. 本节将对零样本分类技术的相关技术体系进行系统梳理. 鉴于现阶段的零样本分类模型的实现都是建立在引入语义辅助信息 (auxiliary information) 的基础之上的. 因此, 本文主要从以下两方面对零样本分类领域的相关技术进行总结, 其一是语义辅助信息, 其二是零样本分类方法.

3.1 用于零样本分类的语义辅助信息

零样本分类需要借助辅助语义信息构建语义空间, 建立视觉模态和语义模态的交互关系. 其中, 辅助语义信息可分为基于人工定义的辅助信息和基于学习的辅助信息两大类. 接下来, 本文将分别从这两大类角度对零样本分类中的辅助信息进行介绍. 此外, 表 1 将辅助信息的分类方式以及各自的优势和不足进行了总结.

表 1 零样本分类中不同种类的辅助信息

Table 1 The different kinds of auxiliary information adopted in zero-shot image classification

Auxiliary information		Advantages	Disadvantages
Human-defined	Attribute based	High accuracy; strong interpretability	High cost for designing attribute; strong subjectivity
	Non-attribute based		
Learning-based	Label embedding based	Free of human annotation; more natural	Weak interpretability; influenced by noise
	Textual embedding based		

3.1.1 人工定义的辅助信息

人工定义的辅助信息是指需要依靠人的领域先验知识, 手动设计辅助信息的每一维度. 使用人工定义的辅助信息构建语义空间的方法, 主要可分为基于属性的方法和非属性方法. 使用属性特征构建语义空间是零样本学习的主流方法之一 [3, 52, 53].

属性是指一个类或对象所具有的某种特性, 其某一维度是对某一类物体例如 “形状、颜色、大小” 等的描述, 是一种更高层次的特征表征. 在属性学习 [40] 领域, 根据属性描述方式的不同, 图像属性大致可分为 3 类: 专家定义的属性 [3]、相对属性 [42]、非人工标注的属性 [43, 44]. 具体到零样本分类任务中, 常用的属性特征分别为二值属性和连续属性. 二值属性是指若该类或对象具有该种属性, 则该维度为 1, 否则为 0; 连续属性是指为进一步表示该类或对象具有该种属性的可能性大小而进行的实值表示.

零样本分类任务中, 非属性的辅助信息则是以构建文本关键词空间的方法为代表. 该类方法首先以类别名称作为索引搜集某类别的文本描述, 然后利用 TF-IDF [54] 等自然语言处理方法提取关键词并将关键词作为语义信息的维度, 主要应用于鸟类 [55]、花类等细粒度图像数据集场景. 某类别的文本描述信息通常来源于维基百科、某一领域的专业网站、搜索引擎或人的主观描述.

总体来看, 人工定义辅助语义信息的优势和不足都比较明显 (见表 1). 其优点在于可充分利用人类在某一领域已有的先验知识设计语义信息的特定维度, 得到的语义向量具有良好的解释性, 判别性好, 准确率高. 同时, 其不足之处在于高度依赖于人工标注, 需要的人力成本和时间成本较高, 且有一定的主观性, 当面对人类先验知识不足的类别时, 难以进行准确标注.

3.1.2 基于学习的辅助信息

基于学习的辅助信息是指利用机器学习等方法, 使用某些预训练的模型, 为每类别学习到一个向量表征作为其语义特征, 此时语义向量的某一维度不再具有实际意义, 该向量整体上代表了该类别的某些特性. 目前, 对于零样本分类任务而言, 基于学习的辅助信息主要可分为标签嵌入方法和文本嵌入方法两个类别.

标签嵌入是指使用 Word2Vec [31] 或 GloVe [49] 等自然语言处理模型将类别标签名称嵌入到一个实值空间, 得到其向量化表示. 利用该方法得到的词向量中向量之间的相似度可以较好地模拟单词语义上的相似度, 有良好的线性相加性, 更加灵活. 且 Word2Vec 或 GloVe 模型可通过维基百科等开源语料库以无监督方式训练得到, 极大地节省了人工标注的成本. 具体来说, 已有部分工作 [56, 57] 利用标签嵌入对语义空间进行构建, 取得了良好的效果.

文本嵌入方法与人工定义辅助信息中的文本关键词方法类似, 首先需要获取关于该类别的文本描述, 但文本嵌入无需挖掘关键词作为向量维度, 而是直接通过某一学习到的文本编码模型将多个文本描述转化为该类别的语义表征向量.

图 3 (网络版彩图) (a) DAP 和 (b) IAP 示意图^[3]. 其中 x 节点表示可观测样本, y 节点表示可见类别, a_m 节点表示属性特征, z_l 节点表示不可见的但需要识别的目标类别.

Figure 3 (Color online) Graphical representation of (a) DAP and (b) IAP^[3]

图 4 各类零样本图像分类方法间的比较

Figure 4 The comparison between different categories of zero-shot image classification approaches

与人工定义的辅助信息不同, 基于学习的辅助信息各维度没有对应的语义表示, 得到的向量表示却也能够模拟类别间的相似度, 以一种更加自然的方式获得语义信息, 其主要优势在于无需人工标注, 可节省大量人力物力. 但是, 相对来说, 该方法得到的语义特征可能包含较多噪声, 且特征可解释性较低, 难以结合人类先验知识. 需要指出的是, 对于现阶段而言, 相较于使用人工定义的辅助信息的零样本分类算法, 在使用基于学习的辅助信息时, 零样本分类模型的性能会产生明显的下滑. 因此, 探索如何提炼出更高效的基于学习的辅助信息对于零样本分类任务具有重要的研究价值.

3.2 零样本分类方法

与其他零样本分类相关的总数文献相比, 本文的主要特点在于以时间为脉络梳理了现阶段零样本分类技术的发展过程. 因此, 我们在此部分中, 仍以时间为线索, 按照不同发展阶段的技术路线, 将现有的零样本分类方法分为 3 类 (如图 3 所示), 即 (1) 基于直接语义预测的方法; (2) 基于嵌入模型的方法; (3) 基于生成模型的方法. 下文将对每种类型的方法进行详细介绍. 此外, 图 4 中对每类方法采用的技术实现框架进行了总结.

3.2.1 基于直接语义预测的方法

早期的零样本分类方法大多数可以归结为基于直接语义预测的方法^[3,52]. 此类方法的主要思路是利用语义属性作为视觉样本的类别标签间桥梁, 通过构建一个基于属性的分类器, 进而对视觉样本的所属类别进行估计.

其中最具有代表性的当属 2013 年 Lampert 等^[3,52] 提出的 DAP (direct attribute prediction) 和 IAP (indirect attribute prediction) 模型. 作为零样本分类的先驱工作, 他们首次利用类别的属性特征计算类别之间的语义关联, 并提出了两种实现零样本分类的概率转移模型, 直接属性预测 (DAP) 和间接属性预测 (IAP). 下面以 DAP^[3] 和 IAP^[3] 为例介绍零样本分类的过程, 如图 3 所示. DAP 是直接学习建立视觉数据与属性特征中每一种属性之间关系的模型, 并利用所学习到的模型预测测试样本的属性特征. IAP 则是利用可见类别将测试样本和未见类别联系起来, 其过程可以分为两步: 首先将测试样本分到可见类别中, 然后利用可见类别和未见类别之间的关系对测试样本进行预测, 其中可见类别和未见类别之间的语义关系是通过类别之间的语义描述建立起来的.

此外, 还有一些工作^[53,58,59] 从属性学习的角度进行零样本分类模型的设计. 其中, Fu 等^[58] 将属性学习的属性空间进行了扩展, 提出了半隐属性 (semilattent attribute) 空间的概念, 即将原本专家定义的确知属性和文中提出的隐属性合并在一个统一的框架下进行表示, 再通过构建一个概率图模型对数据的多模态半隐属性进行属性学习, 最后将样本和标签嵌入到半隐属性空间中进行 k 最近邻 (k -nearest neighbor, kNN) 搜索实现分类. Jayaraman 等^[53] 利用随机森林算法对零样本分类模型进行训练以提高属性预测的准确度, 并在此基础上提出了零样本随机森林模型, 通过对每一个属性的预测错误率进行统计以提升属性预测的准确性. Huang 等^[59] 则使用超图 (hypergraph) 对数据的属性关系进行描述, 通过将视觉特征映射到一个超图嵌入空间实现与属性间的对齐 (aligned), 从而将属性预测问题转化为一个超图分割 (hypergraph cut) 问题进行求解.

使用这种思路所建立的模型具有较强的可解释性. 但是, 其不足之处也较为明显. 例如, 文献^[53] 中就曾指出, 人类对属性的标注并不总是可靠的, 而属性的误标注会对这类方法的性能产生较大负面影响. 此外, 属性间的关联性也会导致冗余信息的产生, 同样会严重影响模型的表现^[60].

3.2.2 基于嵌入模型的方法

为了更好地解决基于直接语义预测的方法中存在的问题, 近几年涌现出许多新的工作, 其中的绝大多数可以归结为基于嵌入模型的方法. 基于嵌入模型方法的核心思想是同时将全部视觉特征和类别标签映射到某一个空间中, 再根据相似性度量进行零样本分类. 依据所嵌入的目标空间的不同, 该方法可以被具体划分为 3 个子类别, 即 (1) 基于语义空间嵌入的方法; (2) 基于视觉空间嵌入的方法; (3) 基于公共空间嵌入的方法.

(1) 基于语义空间嵌入的方法. 该方法^[2,12,61~67] 是将视觉特征和类别标签统一嵌入到语义空间中. 其中, Yu 等^[41] 的思路是从视觉的分布中自动学习一种辨别性的类别描述特征, Gan 等^[63] 利用线性回归模型建立样本与类别之间的关系, 并利用类别名称所代表的语义信息建立类别之间的关系以实现视频中动作的零样本分类. Xian 等^[65] 则认为现有的大多数线性模型并不适用于细粒度图像分类问题, 通过将潜在变量融入兼容性函数中为所有图像-类别对提供了多种线性模型用于选择, 构建了一个具备分段线性性质的非线性映射模型, 再利用基于排序的损失函数进行训练. Fu 等^[66] 考虑到语义空间内部的流形结构保持, 提出使用一种图模型对语义嵌入空间所含的语义流形结构进行建模, 并使用吸收马尔科夫 (Markov) 链在流形结构中进行距离度量.

图 5 (网络版彩图) 各类零样本图像分类方法实现的技术框架

Figure 5 (Color online) The technical frameworks of various categories of zero-shot image classification approaches. (a) Direct semantic predicting based; (b) embedding based; (c) visual data generation based.

随着零样本分类和深度学习技术的进一步发展, 也出现了一些零样本分类方法采用深度网络实现语义空间嵌入. 例如, Frome 等^[45]提出的 DeVISE 模型是较早进行尝试的方法之一. 他们利用 CNN 和 Word2Vec 特征作为输入, 使用合页损失函数构建了一个深度网络结构, 提出了一个能够进行端到端 (end-to-end) 训练的深度零样本分类模型. Socher 等^[12]将已提好的视觉特征通过一个两层或三层的神经网络映射到语义空间, 然后利用最小二乘损失来训练网络.

(2) 基于视觉空间嵌入的方法. 基于语义空间嵌入的方法能够对视觉样本在语义空间中进行表示. 但是, 这种方法较易于受到 hubness 现象的影响, 即在语义空间中存在一些样本会成为许多类别原型的最近邻, 进而导致使用最近邻分类方法进行分类时性能的降低. 因此, 出现了一些方法选择将语义特征嵌入到视觉空间^[61, 62, 68~70]来缓解该问题. 其中, Shigeto 等^[71]从岭回归和最小二乘损失的角度分析了产生 hubness 现象的原因, 并提出了通过转换嵌入方向来实现零样本分类这一任务. Shojaei 等^[70]基于同类样本的视觉特征在视觉空间中聚集在一起的特点, 提出了一种基于聚类的半监督学习方法. 其主要思想是将类别的语义特征映射到视觉特征空间中, 并利用聚类的方法使得类别在视觉空间中的特征之和与类别的语义特征接近. Changpinyo 等^[72]提出了将合成分类器的方法用于零样本分类的方法. 通过引入虚拟类别, 将可见类和不可见类进行连接, 使用同一组虚拟类别对应的分类器作为基分类器构造可见类和不可见类的分类器, 最终基于可见类的数据对合成的分类器进行学习, 并将其应用于不可见类.

(3) 基于公共空间嵌入的方法. 除此之外, 基于嵌入模型的第 3 个子类别的思路是将视觉特征和语义特征嵌入到一个公共空间. 属于该类别的现有工作^[12, 69, 73, 74]大都采用非深度模型进行建模. 例如, Akata 等^[46]利用双线性模型建立视觉样本与类别属性描述之间的兼容度并利用 0-1 损失学习不同类别之间的判别信息. Romera-Paredes 等^[73]利用双线性模型建立视觉样本与类别的属性描述之间的兼容度, 并利用最小二乘回归损失建立兼容度与类别之间的关系, 其算法框架如图 5 所示. 受到字典学习和自训练学习思想的启发, Yu 等^[69]提出了一种直推式零样本图像分类方法, 即先利用字典学习的方法将训练样本的视觉特征和属性特征映射到一个公共的潜在空间进行映射模型的学习, 再使用自训练策略将测试样本融入到模型学习的框架中, 使模型性能得到进一步提升. Guo 等^[74]提出了一种在图像和类别标签间建立直接联系的方法, 在学习类别共享模型空间的基础上, 通过联合学习框架将测试类别也纳入训练阶段用于训练模型.

此外, 也有部分工作将深度网络应用于此类方法当中. 例如, Yang 等^[75]以及 Ba 等^[76]利用图像

和文本描述作为输入, 分别对图像和文本经过若干层的网络变换, 进而在一个公共的空间对它们进行关联分析, 然后利用排序合页损失^[75]或二值交叉熵 (binary cross entropy) 损失^[76]来训练网络. 此外, 相关的研究还有 Reed 等^[50]以及 Zhang 等^[68]的工作. Reed 等^[50]首先通过将图像和相应的文本描述作为输入, 图像部分用 CNN 模型提取特征, 文本描述部分用递归神经网络提取特征, 然后设计了端到端的深度零样本分类模型, 并基于 0-1 损失函数在输出层建立了跨模态的目标函数, 并同时反调两部分网络, 以此建立不同模态之间的语义关联. Zhang 等^[68]的工作与文献 [50] 整体思路较为类似, 不同之处在于他们提出的深度模型能够将文本描述信息和其他语义信息进行融合.

需要指出的是, 基于嵌入模型的零样本分类方法能够在一定程度上缓解前文中提到的 hubness 现象, 但却难以处理零样本分类中固有的另一个难题, 即数据不均衡问题. 因为, 嵌入模型只能够根据已有数据学习得到空间映射函数, 但在处理训练数据较少的类别时, 其训练所得模型的泛化能力较难得到有效保证.

3.2.3 基于视觉样本生成的方法

要解决零样本分类问题中的类别样本不均衡问题, 生成模型是一项有效的技术. 生成模型旨在学习样本数据的概率分布, 这样可以从中随机抽取数据作为数据增强机制. 对于零样本分类任务, 目前使用生成模型的主要应用是生成虚拟视觉样本. 这类方法^[77~85]的主体思路是利用类别对应的语义属性作为控制条件, 生成虚拟的训练数据 (视觉样本), 从而将原本的零样本分类转化为一个传统的全监督分类问题进行处理. 2017 年, Long 等^[78]提出了首个基于未见类样本进行生成的方法. 他们构建了一个能够同时保持跨模态数据分布以及局部信息的语义属性到视觉空间的映射函数, 此外, 还提出了一个称为扩散正则 (diffusion regularisation) 的方法用于保证生成虚拟视觉特征时能够尽量地保证原本的语义结构信息. 此方法在当时取得了最优的表现.

随着深度学习的复兴, 尤其是生成对抗网络 (generative adversarial network, GAN)^[82]这一技术的高速发展, 自 2018 年起, 涌现出了大量利用深度网络进行视觉样本生成的零样本分类方法^[77, 83~85]. 其中的大多数方法是基于生成对抗网络的. 生成对抗网络的主体思想源于博弈论, 它包含一个将噪声作为输入而生成伪数据的生成器, 和一个辨别伪数据和真实数据的辨别器, 经过多次迭代之后, 二者之间达到 Nash 均衡, 生成器生成的伪真实样本能够欺骗判别器, 从而达到生成虚拟真实样本的目的.

针对零样本分类任务, Xian 等^[77]通过生成对抗网络构建了一个生成模型, 判别器部分以类别属性的分类损失进行训练, 生成器部分则借鉴了 Conditional GAN 的思想, 以类别属性为条件生成类别对应的视觉特征. Felix 等^[83]则在文献 [77] 的基础上增加了由视觉空间指向语义空间的回路, 通过构造循环损失函数完成由视觉到语义的对应. 变分自编码器 (variational auto-encoder, VAE) 使用了一个编码器和一个解码器, 该编码器将输入表示为一个符合高斯 (Gauss) 分布的数据, 解码器则由潜在变量重建输入. Shen 等^[84]提出了一个交叉对齐的变分自编码器以确保来自不同输入域的文本的潜在表示形式是相似的. 文献 [85] 则使用一种可比较的方法对来自不同领域的图像的潜在表示进行匹配.

3.3 小结与讨论

经过了近十年的发展, 无论是从辅助语义信息的构建, 还是从零样本分类方法不断演化的角度看, 零样本分类技术都已经取得了令人瞩目的进展. 尤其是得益于深度学习领域的高速发展, 大幅度提升了零样本分类模型的性能.

但从另一角度看, 目前的零样本分类方法仍存在许多不足. 例如, 现有的零样本分类的问题设定实际上与真实世界的情况并不完全符合, 在真实世界中, 我们不可能只要求识别出未见类别, 更多的

图 6 (网络版彩图) AwA 数据集^[3]中带属性标注的图像示例Figure 6 (Color online) Example images in AwA dataset^[3]

需求是既能识别出可见类别又能识别出未见类. 但在这种更真实的设定下, 目前的零样本分类方法仍表现不佳. 此外, 在使用基于学习的辅助语义信息时, 相较于使用人工定义的辅助信息, 目前的零样本分类方法性能会产生大幅下降, 这也阻碍了零样本分类在难以进行人工属性标注的大规模场景下的应用. 因此, 目前的零样本分类领域仍处于高速发展的阶段, 还存在许多亟待解决的问题. 对于未来零样本分类技术发展趋势的论述, 我们会在第 6 节中进行详细讨论.

4 数据集和评价准则

4.1 数据集

在零样本图像分类技术近十年的发展过程中, 涌现出了大量相关的数据集用于性能评测, 这些数据集有的是为零样本图像分类构建, 有的是利用已有的图像分类和细粒度图像分类的数据集, 这些标准的公开数据集能够帮助研究者们对各种算法的相关性能进行有效评价. 本文只对常用的一些数据集进行简要介绍. 常用的图像数据集包括动物类别数据集 AwA^[3] (示例见图 6), 鸟类数据集 CUB^[55], 场景类数据集 SUN^[86], 混合类别 (包含人物、动物、风景等) 的数据集 aPY^[40] 和 ImageNet^[87] 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. 我们还给出了目前在各个数据集上表现最好 (state-of-the-art, SoA) 的算法的性能指标和出处以供参考. 由于目前多个数据集都存在多种数据划分情况, 为了公正评价各算法的性能, 我们给出的算法性能统一以文献 [37] 中提出的数据划分方法作为标准, 统一使用 Resnet101 网络提取视觉特征, 以类平均准确度 (average per-class top-1 accuracy) 作为零样本分类评价指标. 对这些数据集的总结与比较如表 2^[45, 88, 89] 所示. 另外, 我们挑选了 AwA^[3] 数据集中的部分类别以提供一个直观的展示.

表 2 零样本图像分类常用的数据集
Table 2 Popular datasets in zero-shot image classification

Dataset	Numbers of classes	Numbers of instances	Numbers of attributes	Annotation level	SoA
AwA	50	30475	85	Per class	85.3 [88]
aPY	32	15339	64	Per image	39.8 [45]
CUB-200-2011	200	11788	312	Per image	67.8 [88]
SUN-attribute	717	14340	102	Per image	62.4 [88]
ImageNet	22000	15000000	—	Per class	25.4 [89]

4.2 评价准则

通常, 单标签图像的分类精度是根据 Top-1 精度进行衡量的, 对整体数据集则使用平均准确度 (mean average precision, MAP) 进行评价. 但一些数据集存在样本间数量分布不均衡的情况, 如果以 MAP 衡量零样本分类算法性能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, 目前学界广泛采用类平均准确度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

$$M = \frac{\sum_{n=1}^N \text{Acc}_{y_i^{ts}}}{K}, \quad (1)$$

其中 K 表示未见类的总数, y^{ts} 表示未见类, i ($1 \leq i \leq K$) 表示未见类中第 i 个类的分类准确度.

此外, 对于开集零样本图像分类而言, 因为其测试时的分类搜索空间不只局限于未见类别, 还将训练阶段出现的类别也纳入其中, 所以其评价标准也有所不同. 目前尚缺乏对开集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, 类似于 F1 得分 (F1 score), Xian 等 [37] 以算法在训练数据集和测试数据集上分别得到的类平均准确度为基础, 提出了一种名为谐调平均准确度 (harmonic mean accuracy) 的零样本分类性能评价指标, 谐调平均准确度的计算公式为

$$H = 2 \times \frac{\text{Acc}_{y^{tr}} \times \text{Acc}_{y^{ts}}}{\text{Acc}_{y^{tr}} + \text{Acc}_{y^{ts}}}, \quad (2)$$

其中 $\text{Acc}_{y^{tr}}$ 和 $\text{Acc}_{y^{ts}}$ 分别表示在训练数据集和测试数据集上得到的类平均准确度.

5 相关技术的联系与区别

(1) 跨模态学习 (cross-modal learning): 核心在于建立不同模态信息间的关联 [90]. 零样本分类通常需要关联视觉和文本/属性等不同模态间的关系, 因而可以看作是一类跨模态学习技术.

(2) 领域适应学习 [28] (domain adaptation learning, 也称为跨域学习、迁移学习 [26, 32]): 旨在通过学习分布不同但相关的源域和目标域上的数据来获取相应的学习模型, 提高目标域数据分类的精度. 其源域和目标域的数据来自不同领域或者不同视角, 但通常类别相同. 在零样本分类中, 可见类即为源域, 未见类即为目标域, 但两个领域的类别完全不同. 零样本分类可以看作是一类特殊的领域适应学习技术.

(3) 单样本学习 (one/few-shot learning): 利用一个或者极少的标注样本对物体的类别进行学习, 主要利用特征/模型参数共享以及相似文本描述信息等方法加以实现 [10, 11]. 单样本学习属于弱标注情

表 3 零样本分类与相关技术的区别

Table 3 Difference between zero-shot classification and four related techniques

	Cross-modal learning	Domain adaptation	One-shot learning	Anomaly detection	Zero-shot classification
Cross-domain	×	✓	×	✓	✓
Cross-modal	✓	×	×	×	✓
Cross-class	×	×	×	✓	✓

况下的目标分类, 可看作是传统目标分类到零样本分类的一个过渡技术. 此外, 零样本分类的相关方法也可以用于辅助单样本学习.

(4) 异常检测^[91](novelty detection): 旨在判断测试数据是否属于已知类别, 但不测试样本的具体类别做出判断^[92]. 而零样本分类需要对测试样本的具体类别做出判断.

综上所述, 零样本分类可以看作是一类目标域类别没有标注样本的跨模态跨域学习技术. 表 3 中比较了零样本分类技术与 4 种相关技术的区别.

6 面临的问题与挑战

历经十年左右的发展, 零样本图像分类技术发展至今仍面临一些问题和挑战, 大体上可以概括为如下几个主要方面: 高效辅助信息的利用, 多模态映射模型的构建, 映射域偏移 (domain shift), 以及 hubness 现象.

(1) 高效辅助信息的利用. 目前零样本图像分类中常用的辅助信息是属性和词向量. 其中属性通常由手工定义得到, 特别是对于细粒度的数据集标注, 其标注成本十分高昂. 虽然一些基于属性学习的方法相继被提出^[61, 93], 但由于其质量较难保证, 在零样本图像分类中较难获得较高的性能. 这些问题严重限制了属性信息的广泛应用. 对词向量而言, 由于受限于神经语言模型的发展, 因此目前最常使用的仍是 Word2Vec^[31] 和 GloVe^[49] 模型. 相较于属性信息而言, 词向量模型无需手工标注, 因此有着更好的扩展性. 但正是由于这种与领域无关的特性, 使得在零样本图像分类中词向量的性能通常逊色于属性信息. 由于辅助信息是连接可见类别和未见类别, 实现知识有效迁移的桥梁, 因此高效辅助信息的利用是零样本图像分类面临的一个重要挑战. 一些学者也在积极地探索其他辅助信息的利用途径, 特别是网络资源的整合和利用. 例如, 通过维基百科^[46, 76] 和众包技术^[50] 对类别进行准确的文本描述, 或者利用 WordNet^[46] 获得类别的分级信息等.

(2) 域偏移问题. 域偏移问题是领域适应学习中固有的问题. 因为零样本图像分类中同样存在跨域的问题, 因此同样存在域偏移问题. 然而, 与传统的域偏移问题^[26, 32] 不同, 零样本图像分类中的域偏移问题主要是由于训练和测试数据集中所包含的类别不同引起. 因此, 从训练集中学习到的映射函数并不一定适用于测试集. 以 AwA 数据集为例, 图 7(a) 中的斑马和猪的图像是从 AwA 数据集的训练集和测试集中分别随机挑选出的某个类别, 这两个类都具有“有尾巴” (has tail) 这一语义属性, 但从图 7(a) 中可以看出二者“尾巴”的视觉特征完全不同. 图 7(b) 直观地对映射域偏移现象进行了描述, 在 85 维的属性空间中 (为便于可视化, 使用 t-SNE 方法将 85 维的空间图示转化为 2 维平面空间图示), 蓝色数据点分别表示的是斑马 (用 × 表示) 和猪 (用 □ 表示) 这两个类所包含样本的视觉特征经过映射函数映射得到的语义属性映射, 红色五星表示的是斑马和猪这两个类在属性空间中对应的原型 (prototype). 可以明显看出, 使用训练样本 (斑马) 学习到的从视觉空间到属性空间的映射函数, 如果应用于测试类 (猪) 进行映射会导致测试样本和本应与其对应的原型间产生很大的偏差. 现有的技术

图 7 (网络版彩图) 零样本分类中域偏移问题图示

Figure 7 (Color online) An illustration of the domain shift problem in zero-shot image classification. (a) Visual space; (b) attribute space.

主要用两种思路缓解域偏移问题: 一类是基于直推式学习的方法^[47,94,95], 主要思路是通过挖掘未见类样本的流形信息缓解域偏移的程度. 另一类是基于归纳式学习的方法^[66,96], 主要思路是利用来自训练数据的信息对模型施加额外的限制条件. 这些方法能在一定程度上缓解域偏移问题, 但如何进一步缓解或者克服域偏移问题, 在今后很长时间内仍是研究的热点和难点.

(3) Hubness 现象. Hubness 现象是零样本图像分类问题中另一个受到广泛关注的现象, 该问题可以描述为: 在高维空间中对目标进行最近邻搜索时, 会存在一些公共的“邻居”, 它们会干扰并降低预测结果的准确性. Marco 等^[97]首次对该问题进行了研究, 他们认为 hubness 现象是高维向量空间的数据分布中存在的一种固有属性. 为缓解 hubness 现象, Dinu 等^[98]的思路是利用直推式学习, 具体做法是通过对未见类样本的分布进行全局建模. 另外, Shigetou 等^[71]提出将视觉特征空间作为嵌入空间可以有效缓解 hubness 现象, 并进行了相关的理论分析. 值得注意的是, 导致 hubness 现象的原因仍然富有争议. 例如, 与 Marco 等^[97]的观点不同, Low 等^[99]认为 hubness 现象是在数据生成过程中密度变化所引起的一种效应, 他们还通过实验证实了 hubness 不仅会在高维向量空间中出现, 在低维数据空间中同样会存在 hubness 现象. 因此, 可以说有关 hubness 现象的研究仍然是一个开放性课题, 亟待更加深入的研究.

7 发展趋势

本文主要聚焦于介绍零样本图像分类的相关内容. 虽然零样本分类领域已经取得了一些阶段性成果, 但需要指出的是, 目前的零样本分类的研究仍局限于问题设定与真实世界设定相偏离的问题, 想要从实质上推进零样本分类的进展, 必须要认识到目前该领域的不足和局限性. 对此, 我们将主要从 3 个角度出发介绍对未来零样本分类发展趋势的一些思考.

(1) 针对异构样本的零样本分类. 在目前的零样本分类的标准设定下, 可见类(源域)和未见类(目标域)样本的数据类型(特征空间)是相同的, 而二者的标签空间是不同的. 按照迁移学习的定义来说, 迁移学习可以根据源域和目标域的域或任务是否相同, 划分为同构迁移学习(homogeneous transfer learning)和异构迁移学习(heterogeneous transfer learning). 目前绝大多数的零样本分类方法可以看

作是标签空间异构的迁移学习任务. 然而, 在有些场景中, 同数据类型的可见类和未见类样本是难以同时获取的, 此时, 对于可见类和未见类, 只有不同数据类型的样本是可获取的. 例如, 在文献 [100] 的任务中, 可见类的样本是图像数据, 而未见类的样本则是视频数据. 就该任务而言, 相较于传统的监督学习方式, 这种异构样本数据的零样本分类方法能够使训练模型的时间和算力消耗大幅降低. 因此, 我们认为基于异构样本数据的零样本分类在未来会拥有更为广阔的应用前景, 亟待开展进一步研究.

(2) 更加符合真实世界设定的零样本分类. 在目前的零样本分类设定下, 训练集和测试集类别之间是没有交集的, 也就是说, 分类任务仅针对于未见过类的样本. 但是, 实际上这个设定是与我们的常识相违背的, 因为假设所有测试样本全部属于未见类别本身就是过于极端的一种设定. 而在真实场景中, 我们更希望学习到的是不仅能够良好地识别出可见类又能够识别出未见类的模型. 针对这一不足, 目前已经陆续出现一些工作致力于探索更加符合真实世界需求的零样本分类设定. 其中, 目前受到关注较多的是 Chao 等^[101]提出的泛化零样本学习 (generalized zero-shot learning) 设定, 即测试样本既可以来自于可见类别又可以来自于未见类别. 相较于狭义零样本分类, 这一设定显然更加符合真实世界的需求. 相对应地, 近几年中针对泛化零样本分类也出现了一些工作. 如 Arora 等^[89]以变分自编码器 (variational autoencoder, VAE) 为框架构建了一个生成模型, 能够根据类别属性生成对应类别的样本, 从而将零样本分类问题变为了一个传统的全监督分类问题. 类似地, Xian 等^[77]则通过生成对抗网络构建了一个生成模型, 判别器部分以类别属性的分类损失目标进行训练, 生成器部分则是借鉴了 Conditional GAN 的思想, 以类别属性为条件生成类别对应的视觉特征. 需要强调的是, 整体而言, 目前泛化零样本分类模型的性能与狭义零样本分类模型相比还有较明显的差距. 因此, 泛化零样本分类技术的相关研究理应在未来得到更多的关注.

此外, 开集识别 (open set recognition) 也可以看作是更加广泛意义上的一种零样本分类问题. 传统的开集识别概念由 Scheirer 等^[102]在 2013 年左右首次提出, 主要关注的是识别测试样本属于可见类还是未见类, 并不对测试样本的类别做出判断. 此后, Scheirer 等^[103, 104]又将研究课题扩展到了多分类开集识别研究. 与文献 [102] 不同, Bendale 等^[105]在定义开集识别概念时加入了增量学习的思想. 而开集零样本图像分类需要同时对可见类和未见类的数据进行类别判断, 可认为是一种广义的开集识别任务^[56], 这极大增加了研究难度. Socher 等^[12]较早地对开集零样本分类进行了探索, 他们提出利用异常检测的方法将未见类从可见类中区分开来, 并对属于可见类和未见类的样本利用不同的方法进行分类, 这突破了传统的零样本分类中测试样本仅属于未见类别的假设. Zhao 等^[106]提出一种通过将类别标签的流形结构迁移到视觉空间来估计未见类的数据分布的方法. 作为开集识别的一种拓展, Fu 等^[56]提出了泛化开集识别 (generalized open set recognition) 的概念, 在这种设定下, 不仅测试样本不限于来自可见或未见类, 测试阶段的标签还在可见类和未见类之外加入了一些新的类别标签, 这也进一步增大了问题的难度. 针对此任务, Fu 等^[56]提出了一种称为词汇信息学习 (vocabulary-informed learning) 的研究思路, 其主要思想是从一个巨大的词汇表中确定某个图像类别名称, 将词向量 (Word2Vec) 特征作为词汇表, 并利用一种基于最大间隔的语义嵌入框架实现开集零样本分类. 进一步, 他们还与极限值机^[107] (extreme value machine, EVM) 相结合^[108], 利用极限值对视觉特征空间到语义空间的语义嵌入方式进行约束, 同时在语义嵌入空间合成虚拟样本并对训练类别进行极限值编码, 这种方法可以将半监督分类、零样本分类, 以及开集分类问题纳入同一个框架下进行解决.

整体来看, 目前, 更加符合真实场景需求的零样本分类技术依然处于发展的初期阶段, 仍有很多具体问题没有得到解决, 尤其是分类的准确度不够理想, 离现实场景应用仍有较长一段距离.

(3) 与其他学习方式相结合的零样本分类. 从更广泛的意义上来说, 零样本分类可以看作是一种模型学习的方式. 我们可以将其与其他学习方式相结合, 进而应用于其他任务. 例如, 目前已经

出现一些工作将零样本分类技术用于与少样本学习 (few-shot learning) [109,110]、终生学习 (life-long learning) [111] 等技术相结合.

除了与传统的机器学习方式相结合, 作为人工智能领域的另一个研究热点, 将零样本分类与强化学习 [112] 的结合也在不断取得新的进展. 2017 年, Oh 等 [113] 将相似子任务间的对应关系作为强化学习的目标进而设计了一个层次化的深度强化学习模型, 为零样本决策任务提供了泛化能力, 即使得智能体 (agent) 在进行序列决策时能够有效执行在模型训练阶段未曾传授给它的知识 (instructions). Higgins 等 [114] 则针对强化学习中的域适应问题, 基于数据表征与可见域最大化解耦的思想, 设计了一个多阶段强化学习智能体, 实现了零样本条件下的域适应知识迁移. 针对图像解析问题, Liang 等 [115] 提出了一个深度结构可变 (variation-structured) 的强化学习框架, 能够在零样本条件下 (即测试集中的物体间关系未在训练集中出现过) 实现图像中物体间的关系推断. 目前, 将深度学习与强化学习相结合的研究是人工智能领域备受关注的热点课题, 未来也将帮助零样本分类相关领域不断开拓新的边疆.

(4) 更加广泛的应用. 作为零样本学习领域的首个分支任务, 零样本图像分类是本文关注的重点. 但需要指出的是, 现在零样本学习已不仅局限于图像分类. 例如, 2013 年的国际测评 TRECVID 的 MED (multimedia event detection, 多媒体事件检测) 比赛还将零样本事件检测 (0 Ex) 作为一个新的比赛任务 [116]. 该任务只提供相应事件的文本描述而没有视频训练样本, 目的是找到相应的视频, 可以看作是一类视频分类或从文本到视频的跨媒体检索任务 [117,118].

除了图像分类和视频的事件检测, 学者们还将零样本分类应用到图像标注 [15]、多媒体跨模态检索 [16,17]、视频摘要 [18]、情感识别 [19]、行人再识别 [20]、自动问答 [21]、图像风格迁移 [22]、目标检测 [23] 等不同领域, 拓宽了零样本分类的应用范围. 其中, 图像标注和跨媒体检索是比较直接的应用领域. 例如, 在图像标注方面, Zhang 等 [15] 提出了一种基于零样本分类的方法, 克服了图像标注领域中标注内容只能从事先给定的一些标签中选择的不足. 在跨媒体检索方面, 图像分类其实就是给定一个图像, 检索其最近的类别向量的过程; 而从文本检索图像, 实际就是给定一个类别向量, 检索其最近的图像的过程. 因此, 近几年涌现出较多对基于零样本分类的跨媒体检索的工作. 这类工作大多借鉴哈希检索中的思路. 例如, 2016 年, Yang 等 [16] 提出了一种新颖的零样本哈希 (zero shot hashing, ZSH) 方法, 通过利用从可见类样本中学习得到的哈希函数对未见类图像进行编码, 提高了检索的效率. 同年, Guo 等 [17] 提出了一种名为离散相似性迁移网络 (discrete similarity transfer network, SitNet) 的 ZSH 方法, 该方法基于卷积神经网络, 在全连接层之后加入了一层输出二值编码的离散哈希层作为输出层, 同时选择正则化中心损失函数 (regularized center loss) 语义嵌入损失函数 (文中选用最大间隔损失函数) 作为目标函数, 从而将视觉特征嵌入到词向量空间实现了可见类和未见类之间的知识迁移.

此外, 在视频摘要领域, Liu 等 [18] 利用零样本分类技术学习查询词与视频关键帧之间的相似程度, 提升了基于查询的视频摘要的性能. 再如, 在情感识别方面, Xu 等 [19] 提出一种利用文本数据学习情感的零样本情感识别方法, 用于辅助视频的情感识别, 可识别训练过程中未见的情感类型. 2017 年, Wang 等 [22] 提出了一种零样本图像风格迁移方法, 通过联合训练图像迁移网络和参数网络两个 CNN 实现了对未见类图像风格的实时迁移.

8 结束语

大数据助力人工智能达到了一个新的高度. 然而, 基于大数据的人工智能算法还远远不及人类智慧, 因为即使只有少量样本甚至零样本, 人类也具有“触类旁通”或者“照猫画虎”的能力. 零样本图像分类就是试图解决图像理解领域中的此类问题, 经过大约十年的发展, 目前已经取得了长足的进步

与发展. 本文对零样本图像分类的基本原理、关键技术、研究进展、数据和评测、问题与挑战, 以及发展趋势进行了系统性的总结与分析, 希望能够有助于促进该领域的持续发展和进步.

参考文献

- 1 Larochelle H, Erhan D, Bengio Y. Zero-data learning of new tasks. In: Proceedings of AAAI Conference on Artificial Intelligence, Chicago, 2008. 646–651
- 2 Palatucci M, Pomerleau D, Hinton G E. Zero-shot learning with semantic output codes. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2009. 1410–1418
- 3 Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 951–958
- 4 Rohrbach M, Stark M, Schiele B. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Colorado, 2011. 1641–1648
- 5 Habibian A, Mensink T, Snoek C G M. Video2vec embeddings recognize events when examples are scarce. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2089–2103
- 6 Wang R G, Ding K, Yang J, et al. Image classification based on bag of visual words model with triangle constraint. J Softw, 2017, 28: 1847–1861 [汪荣贵, 丁凯, 杨娟, 等. 三角形约束下的词袋模型图像分类方法. 软件学报, 2017, 28: 1847–1861]
- 7 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, 2012. 1097–1105
- 8 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
- 9 Biederman I. Recognition-by-components: a theory of human image understanding. Psychol Rev, 1987, 94: 115–147
- 10 Li F F, Rob F, Pietro P. A Bayesian approach to unsupervised one-shot learning of object categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Madison, 2003. 1134–1141
- 11 Ahsan U, Sun C, Hays J, et al. Complex event recognition from images with few training examples. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, Santa Rosa, 2017. 669–678
- 12 Socher R, Ganjoo M, Bastani H, et al. Zero-shot learning through cross-modal transfer. In: Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, 2013. 935–943
- 13 Ji Z, Yu Y L, Pang Y W, et al. Manifold regularized cross-modal embedding for zero-shot learning. Inf Sci, 2017, 378: 48–58
- 14 Elliott D, Kiela D, Lazaridou A. Multimodal learning and reasoning. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, Berlin, 2016
- 15 Zhang Y, Gong B, Shah M. Fast zero-shot image tagging. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 5985–5994
- 16 Yang Y, Luo Y D, Chen W L, et al. Zero-shot hashing via transferring supervised knowledge. In: Proceedings of ACM International Conference on Multimedia, Amsterdam, 2016. 1286–1295
- 17 Guo Y C, Ding G G, Han J G, et al. SitNet: discrete similarity transfer network for zero-shot hashing. In: Proceedings of International Joint Conference on Artificial Intelligence, Melbourne, 2017. 1767–1773
- 18 Liu W, Mei T, Zhang Y D, et al. Multi-task deep visual-semantic embedding for video thumbnail selection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 3707–3715
- 19 Xu B H, Fu Y W, Jiang Y G, et al. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. IEEE Trans Affect Comput, 2018, 9: 255–270
- 20 Wang Z, Hu R M, Liang C, et al. Zero-shot person re-identification via cross-view consistency. IEEE Trans Multimedia, 2016, 18: 260–272
- 21 Teney D, Hengel A V D. Zero-shot visual question answering. 2016. ArXiv: 1611.05546
- 22 Wang H, Liang X D, Zhang H, et al. ZM-Net: real-time zero-shot image manipulation network. 2017. ArXiv: 1703.07255
- 23 Bansal A, Sikka K, Sharma G, et al. Zero-shot object detection. 2018. ArXiv: 1804.04340
- 24 Dauphin Y N, Tur G, Hakkani-Tür D, et al. Zero-shot learning for semantic utterance classification. In: Proceedings of International Conference on Learning Representations, Banff, 2014

- 25 Johnson M, Schuster M, Le Q V, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *Trans Assoc Comput Linguist*, 2017, 5: 339–351
- 26 Shao L, Zhu F, Li X L. Transfer learning for visual categorization: a survey. *IEEE Trans Neural Netw Learn Syst*, 2015, 26: 1019–1034
- 27 Pan S J, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*, 2010, 22: 1345–1359
- 28 Patel V M, Gopalan R, Li R, et al. Visual domain adaptation: a survey of recent advances. *IEEE Signal Process Mag*, 2015, 32: 53–69
- 29 Deng L, Seltzer M L, Yu D, et al. Binary coding of speech spectrograms using a deep auto-encoder. In: *Proceedings of Annual Conference of the International Speech Communication Association*, Makuhari, 2010. 1692–1695
- 30 Boulanger-Lewandowski N, Bengio Y, Vincent P. Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription. In: *Proceedings of International Conference on Machine Learning*, Edinburgh, 2012
- 31 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of Advances in Neural Information Processing Systems*, Lake Tahoe, 2013. 3111–3119
- 32 Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017. 2962–2971
- 33 Sun Y, Chen Y H, Wang X G, et al. Deep learning face representation by joint identification-verification. In: *Proceedings of Advance in Neural Information Processing Systems*, Montreal, 2014. 1988–1996
- 34 Wu P C, Hoi S C H, Xia H, et al. Online multimodal deep similarity learning with application to image retrieval. In: *Proceedings of ACM International Conference on Multimedia*, Barcelona, 2013. 153–162
- 35 Li H X, Li Y, Porikli F. DeepTrack: learning discriminative feature representations online for robust visual tracking. *IEEE Trans Image Process*, 2016, 25: 1834–1848
- 36 Liong V E, Lu J W, Tan Y P, et al. Deep coupled metric learning for cross-modal matching. *IEEE Trans Multimedia*, 2017, 19: 1234–1244
- 37 Xian Y Q, Schiele B, Akata Z. Zero-shot learning-the good, the bad and the ugly. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017. 3077–3086
- 38 Fu Y W, Xiang T, Jiang Y G, et al. Recent advances in zero-shot recognition: toward data-efficient understanding of visual content. *IEEE Signal Process Mag*, 2018, 35: 112–125
- 39 Wang W, Zheng V W, Yu H, et al. A survey of zero-shot learning. *ACM Trans Intell Syst Technol*, 2019, 10: 1–37
- 40 Farhadi A, Endres I, Hoiem D, et al. Describing objects by their attributes. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 2009. 1778–1785
- 41 Yu F X, Cao L L, Feris R S, et al. Designing category-level attributes for discriminative visual recognition. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 2013. 771–778
- 42 Parikh D, Grauman K. Relative attributes. In: *Proceedings of IEEE International Conference on Computer Vision*, Barcelona, 2011. 503–510
- 43 Alexander S, Forsyth D. Utility data annotation with Amazon Mechanical Turk. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 2008. 1–8
- 44 Liu J G, Kuipers B, Savarese S. Recognizing human actions by attributes. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Colorado, 2011. 3337–3344
- 45 Frome A, Corrado G S, Shlens J, et al. Devise: a deep visual-semantic embedding model. In: *Proceedings of Advances in Neural Information Processing Systems*, Lake Tahoe, 2013. 2121–2129
- 46 Akata Z, Reed S, Walter D, et al. Evaluation of output embeddings for fine-grained image classification. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015. 2927–2936
- 47 Fu Y W, Hospedales T M, Xiang T, et al. Transductive multi-view zero-shot learning. *IEEE Trans Pattern Anal Mach Intell*, 2015, 37: 2332–2345
- 48 Wang X S, Chen C, Cheng Y H, et al. Zero-shot image classification based on deep feature extraction. *IEEE Trans Cogn Dev Syst*, 2018, 10: 432–444
- 49 Pennington J, Socher R, Manning C D. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, 2014. 1532–1543
- 50 Reed S, Akata Z, Honglak L, et al. Learning deep representations of fine-grained visual descriptions. In: *Proceedings*

- of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 49–58
- 51 Kaessli N, Akata Z, Schiele B, et al. Gaze embeddings for zero-shot image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 6412–6421
 - 52 Lampert C H, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans Pattern Anal Mach Intell*, 2014, 36: 453–465
 - 53 Jayaraman D, Kristen G. Zero-shot recognition with unreliable attributes. In: Proceedings of International Conference on Neural Information Processing Systems, Montreal, 2014. 3464–3472
 - 54 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manage*, 1988, 24: 513–523
 - 55 Wah C, Branson S, Welinder P, et al. The Caltech-ucsd birds-200–2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011
 - 56 Fu Y W, Sigal L. Semi-supervised vocabulary-informed learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 5337–5346
 - 57 Lazaridou A, Bruni E, Baroni M. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, Baltimore, 2014. 1403–1414
 - 58 Fu Y W, Hospedales T M, Xiang T, et al. Learning multimodal latent attributes. *IEEE Trans Pattern Anal Mach Intell*, 2014, 36: 303–316
 - 59 Huang S, Elhoseiny M, Elgammal A, et al. Learning hypergraph-regularized attribute predictors. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 409–417
 - 60 Jayaraman D, Sha F, Grauman K. Decorrelating semantic visual attributes by resisting the urge to share. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 1629–1636
 - 61 Akata Z, Perronnin F, Harchaoui Z, et al. Label-embedding for attribute-based classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Portland, 2013. 819–826
 - 62 Norouzi M, Mikolov T, Bengio S, et al. Zero-shot learning by convex combination of semantic embeddings. In: Proceedings of International Conference on Learning Representations, Banff, 2014
 - 63 Gan C, Lin M, Yang Y, et al. Exploring semantic inter-class relationships (SIR) for zero-shot action recognition. In: Proceedings of AAAI Conference on Artificial Intelligence, Austin, 2015. 3769–3775
 - 64 Xu X, Hospedales T, Gong S G. Semantic embedding space for zero-shot action recognition. In: Proceedings of IEEE International Conference on Image Processing, Quebec City, 2015. 63–67
 - 65 Xian Y Q, Akata Z, Sharma G, et al. Latent embeddings for zero-shot classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 69–77
 - 66 Fu Z Y, Xiang T A, Kodirov E, et al. Zero-shot object recognition by semantic manifold distance. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 2635–2644
 - 67 Fu Y W, Yang Y X, Hospedales T, et al. Transductive multi-label zero-shot learning. In: Proceedings of British Machine Vision Association, Swansea, 2015. 37: 2332–2345
 - 68 Zhang L, Xiang T, Gong S G. Learning a deep embedding model for zero-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 3010–3019
 - 69 Yu Y, Ji Z, Li X, et al. Transductive zero-shot learning with a self-training dictionary approach. *IEEE Trans Cybern*, 2018, 48: 2908–2919
 - 70 Shojaei S M, Baghshah M. Semi-supervised zero-shot learning by a clustering-based approach. 2016. ArXiv: 1605.09016
 - 71 Shigeto Y, Suzuki I, Hara K, et al. Ridge regression, hubness, and zero-shot learning. In: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Porto, 2015. 135–151
 - 72 Changpinyo S, Chao W L, Gong B Q, et al. Synthesized classifiers for zero-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 5327–5336
 - 73 Romera-Paredes B, Torr P H S. An embarrassingly simple approach to zero-shot learning. In: Proceedings of International Conference on Machine Learning, Lille, 2015. 2152–2161
 - 74 Guo Y C, Ding G G, Jin X M, et al. Transductive zero-shot recognition via shared model space learning. In: Proceedings of AAAI Conference on Artificial Intelligence, Phoenix, 2016. 3–8
 - 75 Yang Y X, Hospedales T. A unified perspective on multi-domain and multi-task learning. In: Proceedings of Inter-

- national Conference on Learning Representations, San Diego, 2015. 1–9
- 76 Ba J, Swersky K, Fidler S, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 4247–4255
- 77 Xian Y Q, Lorenz T, Schiele B, et al. Feature generating networks for zero-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 5542–5551
- 78 Long Y, Liu L, Shao L, et al. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 1627–1636
- 79 Long Y, Liu L, Shen F, et al. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 2498–2512
- 80 Zhu Y Z, Elhoseiny M, Liu B C, et al. A generative adversarial approach for zero-shot learning from noisy texts. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 1004–1013
- 81 Kumar V V, Arora G, Mishra A, et al. Generalized zero-shot learning via synthesized examples. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 4281–4289
- 82 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems, Montreal, 2014. 2672–2680
- 83 Felix R, Kumar V B G, Reid I, et al. Multi-modal cycle-consistent generalized zero-shot learning. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 21–37
- 84 Shen T X, Lei T, Barzilay R, et al. Style transfer from non-allel text by cross-alignment. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 6830–6841
- 85 Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks. In: Proceedings of Advances in Neural Information Processing Systems, Long Beach, 2017. 700–708
- 86 Patterson G, Hays J. Sun attribute database: discovering, annotating, and recognizing scene attributes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012. 2751–2758
- 87 Rohrbach M, Stark M, Szarvas G, et al. What helps where-and why? Semantic relatedness for knowledge transfer. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 910–917
- 88 Zhang L, Wang P, Liu L, et al. Towards effective deep embedding for zero-shot learning. 2018. ArXiv: 1808.10075
- 89 Arora G D, Verma V K, Mishra A, et al. Generalized zero-shot learning via synthesized examples. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 4281–4289
- 90 Wei Y C, Zhao Y, Lu C Y, et al. Cross-modal retrieval with CNN visual features: a new baseline. IEEE Trans Cybern, 2017, 47: 449–460
- 91 Markou M, Singh S. Novelty detection: a review—part 1: statistical approaches. Signal Process, 2003, 83: 2481–2497
- 92 Zhai S F, Chen Y, Lu W N, et al. Deep structured energy based models for anomaly detection. In: Proceedings of International Conference on Machine Learning, New York City, 2016. 19–24
- 93 Sharmanska V, Quadrianto N, Lampert C. Augmented attribute representations. In: Proceedings of European Conference on Computer Vision, Florence, 2012. 242–255
- 94 Kodirov E, Xiang T, Fu Z Y, et al. Unsupervised domain adaptation for zero-shot learning. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 2452–2460
- 95 Li Y, Wang D H, Hu H H, et al. Zero-shot recognition using dual visual-semantic mapping paths. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 5207–5215
- 96 Changpinyo S, Chao W L, Gong B, et al. Synthesized classifiers for zero-shot learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 5327–5336
- 97 Marco B, Angeliki L, Georgiana D. Hubness and pollution: delving into cross-space mapping for zero-shot learning. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, 2015. 270–280
- 98 Dinu G, Lazaridou A, Baroni M. Improving zero-shot learning by mitigating the hubness problem. In: Proceedings of International Conference on Learning Representations, San Diego, 2015. 1–10
- 99 Low T, Borgelt C, Stober S, et al. The hubness phenomenon: fact or artifact. In: Proceedings of Towards Advanced Data Analysis by Combining Soft Computing and Statistics, Berlin, 2013. 267–278
- 100 Elhoseiny M, Liu J, Cheng H, et al. Zero-shot event detection by multimodal distributional semantic embedding of

- videos. In: Proceedings of AAAI Conference on Artificial Intelligence, Phoenix, 2016. 3478–3486
- 101 Chao W L, Changpinyo S, Gong B Q, et al. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 52–68
- 102 Scheirer W J, de Rezende Rocha A, Sapkota A, et al. Toward open set recognition. *IEEE Trans Pattern Anal Mach Intell*, 2013, 35: 1757–1772
- 103 Scheirer W J, Jain L P, Boulton T E. Probability models for open set recognition. *IEEE Trans Pattern Anal Mach Intell*, 2014, 36: 2317–2324
- 104 Jain L P, Scheirer W J, Boulton T E. Multi-class open set recognition using probability of inclusion. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 393–409
- 105 Bendale A, Boulton T. Towards open set deep networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1563–1572
- 106 Zhao B, Wu B T, Wu T F, et al. Zero-shot learning via revealing data distribution. 2016. ArXiv: 1612.00560
- 107 Rudd E M, Jain L P, Scheirer W J, et al. The extreme value machine. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40: 762–768
- 108 Fu Y W, Dong H Z, Ma Y, et al. Vocabulary-informed extreme value learning. 2017. ArXiv: 1705.09887
- 109 Gan C, Yang Y, Zhu L C, et al. Recognizing an action using its name: a knowledge-based approach. *Int J Comput Vis*, 2016, 120: 61–77
- 110 Tsai Y H H, Huang L K, Salakhutdinov R. Learning robust visual-semantic embeddings. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 3591–3600
- 111 Isele D, Rostami M, Eaton E. Using task features for zero-shot knowledge transfer in lifelong learning. In: Proceedings of International Joint Conference on Artificial Intelligence, New York, 2016. 1620–1626
- 112 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, 518: 529–533
- 113 Oh J, Singh S, Lee H, et al. Zero-shot task generalization with multi-task deep reinforcement learning. In: Proceedings of International Conference on Machine Learning, Sydney, 2017. 2661–2670
- 114 Higgins I, Pal A, Rusu A A, et al. Darla: improving zero-shot transfer in reinforcement learning. In: Proceedings of International Conference on Machine Learning, Sydney, 2017. 1480–1490
- 115 Liang X D, Lee L S Y, Xing E P. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 848–857
- 116 Over P, Fiscus J, Sanders G, et al. TRECVID 2014: an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID, Orlando, 2014. 52
- 117 Chang X J, Yang Y, Hauptmann A G, et al. Semantic concept discovery for large-scale zero-shot event detection. In: Proceedings of International Joint Conference on Artificial Intelligence, Buenos Aires, 2015. 2234–2240
- 118 Xu X, Hospedales T, Gong S G. Transductive zero-shot action recognition by word-vector embedding. *Int J Comput Vis*, 2017, 123: 309–333

A decadal survey of zero-shot image classification

Zhong JI, Haoran WANG, Yunlong YU & Yanwei PANG*

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

* Corresponding author. E-mail: pyw@tju.edu.cn

Abstract Zero-shot image classification refers to learning a visual classifier for categories with zero training examples. This method can effectively solve problems in which the labeled data for some classes are absent and has therefore gained a considerable attention recently. It has been approximately a decade since this technology was first developed. This paper systematically summarizes the research progress over the past decade in this field. First, we introduce the significance and practical application value of zero-shot image classification. Next, the research processes and typical approaches are summarized in detail. Further, we comprehensively review existing datasets and evaluation metrics, together with the relation between zero-shot image classification and other related techniques. Finally, we analyze the hot spots and existing challenges that need to be further studied and emphasize the future trends in this research area.

Keywords zero-shot image classification, attributes, word vectors, cross-modal embedding, domain adaptation learning

Zhong JI received the Ph.D. degree in signal and information processing from Tianjin University, Tianjin, China, in 2008. He is currently an associate professor in the School of Electrical and Information Engineering at Tianjin University. His current research interests include multimedia understanding, zero/few shot learning, cross-modal analysis, and video summarization.

Haoran WANG received his B.S. degree in communication engineering in 2014. He is currently a Ph.D. candidate in the Department of Information and Communication Engineering at Tianjin University, China. His research interests include computer vision and machine learning.

Yunlong YU was born in 1988. He is currently pursuing the Ph.D. degree in the School of Electrical and Information Engineering at Tianjin University. His research interests include machine learning and computer vision.

Yanwei PANG received the Ph.D. degree in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2004. He is currently a professor in the School of Electrical and Information Engineering at Tianjin University, Tianjin, China. His current research interests include object detection and recognition, vision in bad weather, and computer vision.