软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, 2022,33(11):4268–4284 [doi: 10.13328/j.cnki.jos.006336]

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

基于间接域适应特征生成的直推式零样本学习方法*

黄 晟[1,2], 杨万里[1], 张 译[1], 张小洪[1,2], 杨 丹[1,3]

[1](重庆大学 大数据与软件学院, 重庆 401331)

[2](信息物理社会可信服务计算教育部重点实验室(重庆大学), 重庆 400044)

[3](西南交通大学 信息科学与技术学院, 四川 成都 611756)

通信作者: 黄晟 Email: huangsheng@cqu.edu.cn

摘 要: 近年来, 零样本学习备受机器学习和计算机视觉领域的关注. 传统的归纳式零样本学习方法通过建立语义与视觉之间的映射关系, 实现类别之间的知识迁移. 这类方法存在着可见类和未见类之间的映射域漂移(projection domain shift)问题, 直推式零样本学习方法通过在训练阶段引入无标定的未见类数据进行域适应, 能够有效地缓解上述问题并提升零样本学习精度. 然而, 通过实验分析发现, 这种直接在视觉空间同时进行语义映射建立和域适应的直推式零样本学习方法容易陷入"相互制衡"问题, 从而无法充分发挥语义映射和域适应的最佳性能. 针对上述问题, 提出了一种基于间接域适应特征生成(feature generation with indirect domain adaptation, FG-IDA)的直推式零样本学习方法. 该方法通过串行化语义映射和域适应优化过程, 使得直推式零样本学习的这两大核心步骤能够在不同特征空间分别进行最佳优化, 从而激发其潜能提升零样本识别精度. 在 4 个标准数据集(CUB, AWA1, AWA2, SUN)上对 FG-IDA 模型进行了评估, 实验结果表明, FG-IDA 模型不仅展示出了相对其他直推学习方法的优越性, 同时还在 AWA1, AWA2 和 CUB 数据集上取得了当前最优结果(the state-of-the-art performance). 此外还进行了详尽的消融实验, 通过与直接域适应方法进行对比分析, 验证了直推式零样本学习中的"相互制衡"问题以及间接域适应思想的先进性.

关键词: 图像分类; 零样本学习; 生成对抗网络; 域适应; 特征生成

中图法分类号: TP181

中文引用格式: 黄晟, 杨万里, 张译, 张小洪, 杨丹. 基于间接域适应特征生成的直推式零样本学习方法. 软件学报, 2022, 33(11): 4268–4284. http://www.jos.org.cn/1000-9825/6336.htm

英文引用格式: Huang S, Yang WL, Zhang Y, Zhang XH, Yang D. Feature Generation Approach with Indirect Domain Adaptation for Transductive Zero-shot Learning. Ruan Jian Xue Bao/Journal of Software, 2022, 33(11): 4268–4284 (in Chinese). http://www.jos.org.cn/1000-9825/6336.htm

Feature Generation Approach with Indirect Domain Adaptation for Transductive Zero-shot Learning

HUANG Sheng[1,2], YANG Wan-Li[1], ZHANG Yi[1], ZHANG Xiao-Hong[1,2], YANG Dan[1,3]

[1](School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China)

[2](Key Laboratory of Dependable Service Computing in Cyber Physical Society (Chongqing University), Ministry of Education, Chongqing 400044, China)

[3](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: In recent years, zero-shot learning has attracted extensive attention in machine learning and computer vision. The conventional inductive zero-shot learning attempts to establish the mappings between semantic and visual features for transferring the knowledge

* 基金项目: 国家重点研发计划(2018YFB2101200); 国家自然科学基金(61772093, 61602068); 中央高校基本科研业务费专项资金(2019CDYGYB014)

收稿时间: 2020-12-16; 修改时间: 2021-01-25; 采用时间: 2021-03-12

between classes. However, such approaches suffer from the projection domain shift between the seen and unseen classes. The transductive zero-shot learning is proposed to alleviate this issue by leveraging the unlabeled unseen data for domain adaptation in the training stage. Unfortunately, empirically study finds that these transductive zero-shot learning approaches, which optimize the semantic mapping and domain adaption in visual feature space simultaneously, are easy to trap in “mutual restriction”, and thereby limit the potentials of both these two steps. In order to address the aforementioned issue, a novel transductive zero-shot learning approach named feature generation with indirect domain adaption (FG-IDA) is proposed, that conducts the semantic mapping and domain adaption orderly and optimizes these two steps in different spaces separately for inspiring their performance potentials and further improving the zero-shot recognition accuracy. FG-IDA is evaluated on four benchmarks, namely CUB, AWA1, AWA2, and SUN. The experimental results demonstrate the superiority of the proposed method over other transductive zero-shot learning approaches, and also show that FG-IDA achieves the state-of-the-art performances on CUB, AWA1, and AWA2 datasets. Moreover, the detailed ablation analysis is conducted and the results empirically prove the existence of the “mutual restriction” effect in direct domain adaption-based transductive zero-shot learning approaches and the effectiveness of the indirect domain adaption idea.

Key words: image classification; zero-shot learning; generative adversarial network; domain adaptation; feature generation

在图像领域传统的分类方法中, 训练分类器需要大量的图片样本和对应的标签. 但现实场景下, 某些类别的样本可能是十分稀少甚至缺乏的, 如濒临灭绝的物种鸭嘴兽, 由于其数量稀少且行踪隐秘, 通常无法收集足够的图片样本以支撑该物种分类器的有效训练. 针对这一问题, Lampert 等人在 2009 年首次提出了零样本识别问题, 旨在解决现实场景下缺乏标定数据的物体识别问题, 引发了学术界对零样本学习的广泛关注[1].

Lampert 等人提出的零样本学习的核心思想是, 借助类别描述来解决样本缺失的问题. 在 Lampert 提出的 Animal with Attributes (简称 AWA)数据集中, 分为可见类和未见类两种样本集, 前者具有充足的动物标定图像可以利用, 而后者没有任何用于训练的图像. 在训练阶段提取可见类的分类知识, 并构建可见类与未见类之间的联系, 将可见类上的分类知识迁移到未见类上, 从而在测试阶段能够对未见类进行分类. 例如, 一个从来未曾见过鸭嘴兽但见过其他动物的人, 在“鸭嘴兽同时有鸭子和哺乳动物特点”描述的帮助下, 结合所见过的动物信息, 在碰见鸭嘴兽时能够识别出该物种. 这就是零样本学习识别未见类的过程[1,2].

现有的零样本学习方法从类别描述的语义信息入手, 通过可见类与未见类之间的语义联系, 完成知识迁移[3]. 目前, 传统的语义信息有属性(attribute)和词向量(Word2Vector)两种. 属性通常是人工标定的类别描述, 如 AWA 数据集中的“外形”“花纹”“颜色”等, 通过设定属性向量的各维度数值来描述相关类别. 而词向量则是从人类的知识语料库(如维基百科)中提取出类别的释义, 然后通过自然语言处理技术转换为词向量. 目前的零样本学习方法更多的是使用属性作为语义信息.

早期的零样本学习论文中[4−7], 通过将视觉信息映射到语义空间中进行分类. 在可见类数据上训练的主要目的是构建一个从视觉信息到语义空间的良好映射函数, 使得在语义空间中能够通过最近邻方法正确完成分类任务, 随后将该映射函数迁移至未见类的视觉空间中进行映射分类. 在后续的研究中, 也有将语义信息投影到视觉空间[8]以及将语义信息和视觉信息共同投影到中间空间的方法[9]. 而近年来, 由于直接解决了未见类样本缺失的问题, 基于生成模型的方法逐渐成为零样本学习领域的主流. 这些基于生成模型的方法通过在可见类上学习到的语义映射, 利用未见类的语义表征生成未见类的视觉特征, 从而将零样本学习问题转化为普通分类问题予以解决. 在零样本学习领域, 基于生成模型的方法无论是在稳定性还是识别精度上, 都取得了长足的进步.

近年来, Xian 与 Fu 等人[3,10]指出: 由于零样本问题中的未见类视觉信息缺失, 导致可见类上语义和视觉之间的知识在迁移至未见类时, 会难以避免的产生偏差. 例如属性“有尾巴”和尾巴视觉特征之间的映射关系: 由于可见类(如老虎、野马、水獭的尾巴)与未见类(如猪的尾巴)在视觉特征上有较大差异, 且未见类的视觉特征缺失, 如果直接把可见类上建立的“有尾巴”的属性和对应视觉特征之间的映射关系迁移至未见类上, 那么未见类如猪的尾巴视觉特征将无法良好地反映“有尾巴”的属性, 导致视觉与属性无法正确对齐. Xian 与 Fu 等人在研究中将这个问题称为 Projection Domain Shift, 即映射域漂移.

由于零样本学习问题设置的特殊性, 映射域漂移成为零样本学习不可避免的问题. 针对这一问题, 部分

学者开展了直推式零样本学习研究. 该方法在训练阶段采用直推式数据设置而非传统的数据设置[11,12]. 直推式的数据设置在训练阶段允许使用未见类的无标定视觉特征, 以用于调节可见类与未见类之间的映射域漂移引起的误差. 近年来, 诸多直推式零样本学习取得了令人瞩目的性能, 相对于传统零样本学习方法性能优势明显.尤其是在特征生成方法中, 部分直推式特征生成模型已达到当时零样本领域的最高精度[13,14].

直推式生成模型 ADA[13]等, 本质上将未见类的生成特征修正视为一种特征域适应方法. 因此, 生成模型首先利用已标定的可见类数据学习到“语义→视觉特征”映射模型, 并通过可见类和未见类的语义关系生成相应视觉特征, 即“特征生成”. 然而, 在可见类上学习的映射模型生成的视觉特征必然会对可见类产生偏置. 因此, 直推式生成模型引入未见类的真实特征直接在原空间下对生成模型构建的映射关系进行修正, 从而减轻生成模型对可见类的偏置, 生成更优质真实的未见类视觉特征. 在 ADA[13]的工作中, 将该过程被称为未见类的特征域适应(以下简称域适应). 因此, 直推式生成方法通常完成两个任务: “语义→视觉特征”映射模型的建立和未见类生成特征的域适应优化.

目前, 大多数基于特征生成的直推式零样本学习模型都是在视觉空间中, 通过直接优化生成关系完成域适应, 如 F-VAEGAN-D2[8], ADA[13]等. 本文将其称为直接域适应特征生成模型(feature generation with direct domain adaption, FG-DDA). 通过对近几年的 FG-DDA 模型进行研究分析, 发现这类模型对未见类视觉信息的利用过于简单粗糙. FG-DDA 模型在进行特征域适应时, 通常直接将未见类视觉特征引入原特征空间下, 对原模型进行调整. 这种简单的域适应思路会造成“语义→视觉特征”映射与未见类生成特征的域适应优化过程产生“相互制衡”效应, 而无法同时给出最优的“语义→视觉特征”映射模型和未见类数据的域适应模型. 具体来说, 由于“语义→视觉特征”映射模型是通过对“内容”填充“细节”从而生成高质量样本, 而域适应则是使用无监督的方式对未见类的生成细节作修正, 显然, 在同一个空间内同时优化这两个任务往往会导致两者相互冲突. 以狐狸类的“耳朵”属性为例子, 在“语义→视觉特征”映射阶段即通过语义生成数据的过程中, 需要补充“绒毛”“质地”“形状”等细节信息来生成更为真实的狐狸耳朵相关特征. 但往往生成模型仅保存了可见类中其他动物的相关耳朵信息, 导致生成的耳朵信息向可见类偏置. 而域适应任务则是要求生成特征“耳朵”需符合真实特征集中“耳朵”的共性视觉特征, 剥离含有域信息的不必要细节特征. 总的来说, FG-DDA 方法将映射学习和域适应任务放置在视觉特征空间下同时进行, 一方面要求生成符合可见类规则的未见类特征, 一方面又要求生成特征符合未见类数据特征, 这种并行完成两项任务的做法无法使语义映射的建立和特征域适应任务均完美构建.

如图 1 所示, 特征生成框架在迁移可见类上建立的语义映射到未见类时, 由于映射域漂移问题, 使未见类生成样本(虚线黑点)与真实样本(实线红点)存在分布差异. 本文针对上述问题, 通过引入低维嵌入空间进行未见类样本的间接域适应, 解决原特征空间下存在的分布差异. 本文提出的思路避免了传统 FG-DDA 模型将“语义→视觉特征”映射学习与未见类特征域适应任务在同一个特征空间同时进行优化, 缓解了两者之间的“相互制衡”效应. 根据模型特点, 本文称该特征生成模型为间接域适应特征生成模型(feature generation with indirect domain adaption, FG-IDA). 如图 2 所示, FG-IDA 模型在传统条件生成对抗网络(conditional generative adversarial networks, C-GAN)基础上额外添加了间接域适应模块. 该模块首先将未见类的真实特征与生成特征投影到一个低维嵌入空间, 在该空间中, 利用一个域分类网络对未见类的真实特征与生成特征进行域分类, 同时还利用一个分类网络对生成特征低维嵌入特征进行分类, 同时确保了对嵌入特征的正确分类以及生成特征与真实特征嵌入空间下的域相容, 从而完成间接域适应任务. FG-IDA 模型不仅弱化了域适应与映射学习之间的“相互制衡”, 同时还对视觉特征进行二次特征提取, 通过间接域适应方式提取与未见类物体内容相关的视觉特征, 同时去除与可见类细节相关且干扰未见类分类识别的视觉特征, 规避“语义→视觉特征”映射学习过程中产生的语义鸿沟问题, 提供一个更具判别能力的紧致数据表达.

本文的主要贡献如下: (1) 首次介绍了传统直推式零样本学习框架中在同一空间中进行语义映射和域适应学习的“相互制衡”问题, 并运用实验分析对其予以论证; (2) 提出一种称为间接域适应特征生成模型(FG-IDA)的全新直推式零样本学习框架, 该方法通过引入低维嵌入空间进行间接特征域适应, 缓解了“相互制衡”

问题, 从而更好地解决基于特征生成的直推式零样本学习模型中的“语义→视觉特征”映射及未见类数据的域适应问题; (3) 本文实现了 FG-IDA 模型的端到端(end-to-end)学习, 通过在间接域适应模块中引入分类网络, 能够直接对样本进行分类, 无须专门离线训练分类器; (4) FG-IDA 模型提供了一种简洁的低维特征学习框架, 能够学习一种对未见类生成特征与真实特征均通用且具备良好判别能力的紧致视觉特征; (5) 4 个常见标准数据集上的大量实验结果表明, FG-IDA 模型具备优异零样本学习性能, 特别是在 AWA1, AWA2 以及 CUB 数据集上, 它已取得了当前最优的识别精度.

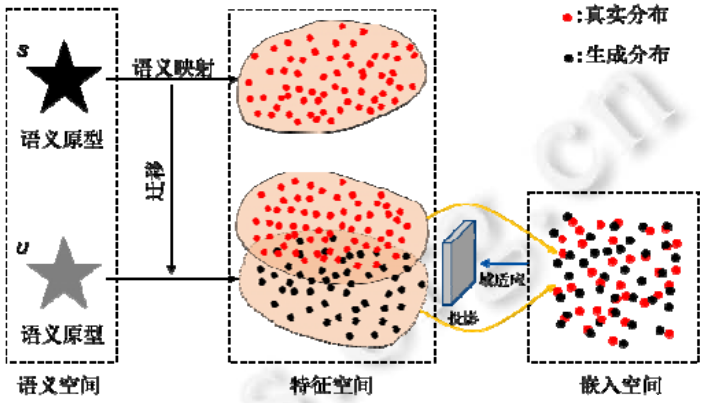
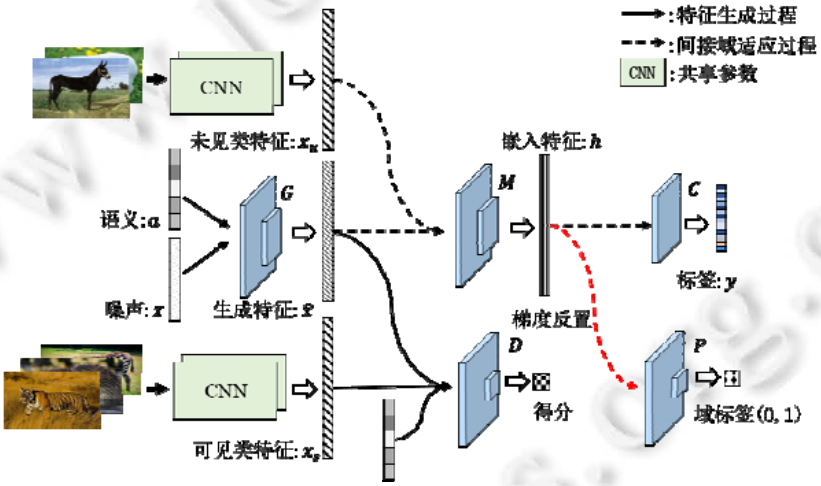


图 1 基于间接域适应的“相互制衡”效应主要解决思路示意图



G: 生成器 *D*: 判别器 *M*: 投影网络 *C*: 分类网络 *P*: 域分类网络

图 2 FG-IDA 模型算法思路图

1 相关工作

零样本学习是机器学习中的一个热点问题, 在问题设置上, 与传统分类任务的关键区别是: 在测试时, 模型会对训练期间未见过的类别样本进行分类. 这个问题在计算机视觉、自然语言处理和机器感知中得到了广泛的研究[3]. 在常规的机器学习中, 分类器应将新样本正确分类为在训练期间已经训练过的类别; 而在零样本学习中, 分类器在训练期间并不会接收来自测试类别的样本.

早期的零样本学习[6,15−19]利用一种双阶段(two-stage)策略来推测图像所属的未见类标签. 通常来说, 就是在第 1 阶段, 输入图像预测其属性; 然后在第 2 阶段, 通过搜索与其最相似的属性来预测该图像所属的类别标

签. 例如在 DAP[15]中, 首先通过学习一个属性分类器来预估图像的每个属性概率; 然后计算类别后验, 并使用 MAP 评估来预测类别标签. 同样的, 也有方法首先为每个属性学习一个概率分类器, 然后通过能够处理未见类属性的随机森林来预估图像的类别概率[16]. IAP 首先预测可见类的类别后验, 然后用每个类别的概率来计算图像的属性后验. 其中, 可见类的类别后验是由多分类器预测的. 此外, 这种 two-stage 方法还能扩展到属性未见的情况. 例如, IAP, CONSE[17]首先预测可见类的后验, 然后通过采用 Top-T 预测类别的组合, 将图像特征投影到 Word2Vec[20]空间中. Two-stage 方法的痛点在于中间任务和目标任务之间存在的映射域漂移[7].例如, 零样本的最终任务是预测标签, 但 DAP 却着重于中间任务-属性分类器的学习, 而中间任务和最终任务的漂移却没有顾及. 除了将视觉空间向语义空间进行映射分类的方法之外, 也有人提出了将两者投影到共同空间下进行分类的思路. Schonfeld 等人[9]提出了 CADA-VAE 模型, 通过两个并列的变分自编码, 将视觉空间和语义空间共同投影到子空间下进行对齐分类.

在后来的工作中, 生成模型[21−24]的引入使零样本学习的准确率结果有了极大的提升. 生成方法除了精度优异, 还能通过使用合成样本, 将零样本学习问题转换为可以处理可见类偏置的常规监督学习问题. 在生成方法流行的过程中, 很多生成模型被提出: Verma 等人[23]使用了基于指数簇框架的简单生成模型, Li 等人[25]通过优化真实分布与生成分布之间的最大平均差异(MMD)[26]得到生成样本. 目前最流行的生成方法是采用 VAE[27]和 GAN[28]的深度生成模型进行样本生成, 如, Mishra 等人[29−31]采用基于 VAE 的体系结构, 而 Lu 等人[21,22,32]则是采用带属性条件的生成对抗网络. 在所有生成方法中, 基于 VAE 或 GAN 的深度生成模型往往能够取得优异效果. 然而与上述其他零样本学习方法相似, 大部分生成模型仍无法有效地缓解映射域漂移问题.

在零样本学习发展前期, 很少有研究者关注映射域漂移问题, 直到直推式零样本学习方法被提出. Verma 等人[23]通过简单的高斯混合模型更新对未见类进行域适应. Song 等人[33]在直推式设置中使用了无偏置嵌入进行了未见类的域适应. Kodirov 等人[34,35]提出了零样本识别问题的无监督域自适应[36]. Zhang 等人[37]使用结构化 SVM 公式进行域自适应. 在生成方法流行起来后, 直推式生成方法由于其优越的精度逐渐成为主流. Xian 等人[8]使用条件 GAN 和 VAE 相结合的生成模型, 并在判别器中并行加入了一个无条件判别器, 用于使用未见类的视觉特征进行域适应. Khare 等人[13]对生成模型采用了 ADA (adversarial domain adaptation)的域适应方法, 利用对抗思想实现了源域到目标域以及目标域到源域的域适应. 以上方法均是在视觉特征空间下构建域适应任务, 通过引入未标定的未见类视觉特征, 在建立语义到视觉的映射同时进行直接的偏置修正. 本文将这类域适应方法统称为直接域适应方法, 也就是 FG-DDA 模型. 然而, 由于语义到视觉特征映射学习与特征域适应在目标优化上存在着一定程度的冲突, 该方法不可避免地陷入了两项任务之间的“相互制衡”问题, 从而限制了模型的整体性能. 本文提出一种间接域适应框架, 通过引入一个间接域适应空间, 使得上述任务在不同的特征空间进行优化, 调和任务之间冲突, 缓解“相互制衡”效应.

2 基于间接域适应特征生成的直推式零样本学习方法

在零样本学习问题中, 数据可以分为可见类 S 与未见类 U 两类数据, 它们由$\{X_s,A_s,Y_s\}$/$\{X_u,A_u,Y_u\}$表示, 分别代表视觉特征、语义和标签. 本文在语义信息上采用的是类别属性 A, 其中,

- $X_s=[x_1^s,...,x_m^s]\in R^{d\times m}$, $X_u=[x_1^u,...,x_n^u]\in R^{d\times n}$ 分别表示来自 S/U 域的 m/n 个 d 维特征向量;
- $A_s=[a_1^s,...,a_{K_s}^s]\in R^{o\times K_s}$, $A_u=[a_1^u,...,a_{K_u}^u]\in R^{o\times K_u}$ 分别表示来自 S/U 域的 k_s/k_u 个 o 维类别语义向量;
- $Y_s\in R^{1\times m}$, $Y_u\in R^{1\times n}$ 分别来自 S/U 域各类别的标签.

在零样本学习设置中, 可见类与未见类两个域的数据没有交集, 即 $Y_s\cap Y_u=\varnothing$. 属性集为 $A=A_s\cup A_u$, 属性向量的数量为 $k=k_s+k_u$, 样本集 $X=X_s\cup X_u$. 零样本学习需要完成的任务是: 在训练阶段, 借助 X_s, A_s, Y_s 和 A_u 训练出模型, 在测试时能够对 X_u 进行正确分类.

由于 $Y_s\cap Y_u=\varnothing$, 在视觉空间中学习到的可见类的分类知识无法应用到未见类的识别任务中. 因此, 对未见类的识别只能借助类别属性作为知识迁移的媒介进行实现. 在可见类上学习 X_s 和 A_s 的映射关系, 通过 A_s 与 A_u 之间的语义关联关系, 将可见类上所学习的知识迁移至未见类上, 最终完成对 X_u 的分类.

对于本文采用的生成方法, 大致思路为: 在可见类上训练得到一个从语义到视觉特征的映射 $A_s \rightarrow X_s$, 然后使用未见类的属性集合 A_u 作为输入, 得到未见类的生成样本集, 并在该样本集上直接训练分类器, 实现对 X_u 的分类. 而直推式零样本学习问题, 还需要使用未见类的未标定视觉特征进行生成特征的域适应, 缓解映射域漂移问题. 但传统的直推式生成方法通常是一个 FG-DDA 模型, 即映射建立任务与未见类特征域适应任务在原特征空间下同时优化, 这导致了前文所提到的"相互制衡"效应.

本文提出了间接域适应特征生成模型(feature generation with indirect domain adaption, FG-IDA), 以解决直推式零样本问题. 如图 2 所示, FG-IDA 模型具有两个核心模块: 一个特征生成模块, 一个间接域适应模块, 分别对应基于特征生成的直推式零样本学习中的两个核心任务: "语义→视觉特征"映射模型的建立和未见类生成特征的域适应优化.

2.1 特征生成模块

此小节将介绍 FG-IDA 模型的生成模块, 该模块的主要任务是, 在可见类上构建语义向视觉特征的映射 $A_s \rightarrow X_s$.

如图 2 所示, FG-IDA 引用了 Conditional Generative Adversarial Network(简称 CGAN)[38]进行特征生成. GAN 网络通过对抗训练, 同时提升判别器的判别能力和生成器的生成能力, 最终得到能够生成逼真样本的生成器网络. 而 CGAN 可在传统生成对抗网络的基础上实现条件生成, 通过输入条件得到条件对应的视觉特征. CGAN 在本质上实现的是一个从语义条件到视觉特征的映射, 即 $A_s \rightarrow X_s$.

CGAN 包含生成器网络和判别器网络: 生成器 $G(z,a;\theta_g)$使用噪声和语义作为输入, 输出语义对应的生成特征; 判别器 $D(x,a;\theta_d)$的输入为生成特征或真实特征以及相关语义, 输出关于特征真假性的对应得分. 在训练过程中, 生成器 G 与判别器 D 迭代更新: 首先, 将真实样本 x 和生成样本 $\hat{x}$ 以及对应属性 a 输入判别器 D, 分别向高分和低分优化, 从而训练判别器 D 根据语义判断输入特征真实性的能力; 随后, 固定判别器 D, 将生成器 G 得到的生成特征 $\hat{x}$ 和对应属性 a 输入判别器 D, 利用 D 输出的得分优化生成器 G, 使其能生成更加真实的特征. 以上过程在训练中反复迭代进行, 逐步提升生成网络和判别网络的性能, 使得最终得到的生成器 G 能够拟合可见类上语义到视觉特征的映射 $A_s \rightarrow X_s$.

CGAN 的训练过程可看作以下最大最小优化问题:

$$\min_G \max_D \mathcal{L}_{CGAN} \coloneqq E[\log D(x,a)] + E[\log(1-D(\hat{x},a))] \tag{1}$$

其中, $\hat{x} = G(z,a)$. 虽然生成对抗网络已经被证明可以拟合复杂的数据分布, 但由于其对抗过程复杂导致难以训练. 为了改进该问题, 本文采用了利于训练的 WGAN-GP 优化版本[39]. 该版本为判别器网络 D 新增一项梯度惩罚的损失项, 新的生成对抗网络损失为

$$\mathcal{L}_{C\text{-}WGAN\text{-}GP} = E[D(x,a)] - E[D(\hat{x},a)] - \lambda E[(\|\nabla_{\tilde{x}} D(\tilde{x},a)\|_2 - 1)^2] \tag{2}$$

损失$\mathcal{L}_{C\text{-}WGAN\text{-}GP}$的前两项代表生成对抗过程: WGAN-GP 不使用对数做损失, 判别器直接对真样本的正得分和假样本的负得分正向优化即可. 最后一项是判别器的梯度惩罚, 其中, $\tilde{x} = \alpha x + (1-\alpha)\hat{x}$, $\alpha \sim U(0,1)$; λ是一个正值惩罚权重, 该惩罚项保证在判别器 $G(z,a)$向真实分布靠齐时, $D(G(z,a),a)$不会超过 $D(x,a)$, 避免优化过程崩溃, 使生成对抗网络更易于训练.

FG-IDA 模型的特征生成模块为 CGAN, 使用了 WGAN-GP 优化版本, 简称 C-WGAN-GP. 它通过最大最小博弈训练, 实现了生成器 G 对可见类的真实分布的拟合, 得到了语义到视觉特征映射: $A_s \rightarrow X_s$.

2.2 间接域适应模块

如前文所述, 通过可见类信息训练出的生成器 G 无法良好地反映未见类中存在的语义映射关系. 如果直接将该映射迁移至未见类, 会不可避免地产生映射域漂移, 导致未见类的生成特征与真实特征存在分布差异. 本节将介绍 FG-IDA 如何通过间接域适应缓解该问题.

如图 2 所示, FG-IDA 的间接域适应模块由 3 个网络构成: 投影网络 $M(x;\theta_m)$, 通过输入特征空间下的真实特征或生成特征, 输出对应特征在 q 维嵌入空间下的投影, 本文将其称为嵌入特征; 类别分类网络 $C(h;\theta_c)$, 通

过输入嵌入特征, 输出该嵌入特征对应的类别标签; 域分类网络 $P(h;\theta_m)$, 通过输入嵌入特征, 输出嵌入特征所对应的域标签.

在特征空间下, 存在真实未见类特征分布 X_u 和生成未见类分布 $\hat{X}_u$. 首先, 利用网络 M 对 X_u 和 $\hat{X}_u$ 进行投影, 得到嵌入空间下的真实嵌入分布 H 和生成嵌入分布 $\hat{H}$; 随后, 由域分类网络 P 对 H 和 $\hat{x}$ 分布进行域分类, 捕捉分布差异信息; 同时, 使用类别分类网络 C 在生成嵌入分布 $\hat{H}$ 上建立分类任务, 再将类别分类网络 C 的分类信息以及域分类网络 P 的分布差异信息均回传给投影网络 M. 不同的是, 前者正向回传, 后者则反向回传. 这使得投影网络 M 能够在保证分类任务准确的同时, 通过减小分布差异的方式实现 H 和 $\hat{H}$ 的分布对齐. 并且根据嵌入空间下分布对齐的特性, 还可以把在生成嵌入分布 $\hat{H}$ 上学习到的样本分类知识迁移至真实嵌入分布 H, 从而实现模型对未见类真实样本的直接分类.

类别分类网络 C 的作用是, 在 $\hat{H}$ 分布上建立分类任务. 由于生成嵌入分布 $\hat{H}$ 是由未见类生成特征分布 $\hat{X}_u$ 投影得到, 因此拥有标签信息. 将生成嵌入特征 $\hat{h}\sim\hat{H}$ 输入类别分类网络 C, 可获得 k_u 维类别预测输出 $\tilde{y}$, 并与对应标签向量 y 做交叉熵以度量分类损失:

$$\mathcal{L}_C=-\sum_{i=1}^{K_u}y_i\times\log\tilde{y}_i \tag{3}$$

域分类网络 P 的作用是区分 H 和 $\hat{H}$ 分布, 保存两者的分布差异信息. 其中, H 的域标签 l=0, $\hat{H}$ 的域标签 l=1. 同样地, 将嵌入特征输入域分类网络 P 中, 获得域分类的预测输出 $\tilde{l}$, 并与其真实域标签 l 做二分类交叉熵损失:

$$\mathcal{L}_P=-[l\times\log\tilde{l}+(1-l)\times\log(1-\tilde{l})] \tag{4}$$

投影网络 $\hat{H}$ 的作用是, 提取特征空间下 X_u 和 $\hat{X}_u$ 的嵌入特征并实现间接域适应. 该任务不仅要求 $\hat{H}$ 和 H 的分布对齐, 同时也需要保证 $\hat{H}$ 上建立的分类任务能够准确地迁移至 H. 因此, 投影网络 M 需要最小化公式(3)中的类别分类损失, 同时最大化公式(4)中的域分类损失. 这需要域分类损失$\mathcal{L}_P$ 对域分类网络 P 直接优化, 对投影网络 M 反向优化. FG-IDA 通过在 P 与 M 之间增加一层梯度转置层实现这一目的. 梯度转置层保证由 M 得到的嵌入特征 h 准确而完整地前向传入域分类网络 P. 但域分类的交叉熵损失$\mathcal{L}_P$回传给 P 后, 在继续反向传播至 M 时, 需要将梯度乘以转置系数$-\beta$(其中, $0<\beta<1$), 使用$\mathcal{L}_P$损失更新域分类网络 P, 使用$-\beta\mathcal{L}_P$损失更新投影网络 M.

综上所述, 间接域适应模块优化模型定义如下:

$$\min_{M,C,P}\mathcal{L}_M:=\mathcal{L}_C-\beta\mathcal{L}_P \tag{5}$$

损失$\mathcal{L}_M$的第 1 项代表间接域适应模块建立的嵌入特征分类任务, 第 2 项代表嵌入特征的分布对齐. 值得注意的是: 损失第 2 项中的系数$-\beta$由梯度转置层实现, 域分类网络 P 采用的损失为公式(4).

通过优化上述模型, 投影网络 M 能够对嵌入特征进行分布对齐, 使得原特征空间下未见类生成特征与真实特征存在的差异, 在嵌入空间下得到了弥补. 这个过程间接地实现了 X_u 和 $\hat{X}_u$ 的域适应, 缓解了映射域漂移问题.

FG-IDA 模型的间接域适应模块在原有视觉特征基础上提取一种对域信息更不敏感的紧致特征, 并以该特征为桥梁, 将生成样本上的分类知识迁移到真实样本上, 实现了在嵌入空间下对真实样本的直接分类. 并且由于嵌入空间的引入, 间接域适应模块能够独立于特征生成模块运行, 有效地缓解两大核心步骤在同一特征空间下同时优化所引起的“相互制衡”效应.

2.3 FG-IDA模型

如上文所述, FG-IDA 模型由特征生成模块和间接域适应模块组成. 通过如图 2 所示的网络结构串联特征生成模块与间接域适应模块, 得到端到端(end to end)训练的 FG-IDA 模型:

$$\min_{G,M,P,C}\max_{D}\mathcal{L}_{FG\text{-}IDA}=\mathcal{L}_{C\text{-}WGAN\text{-}GP}+\gamma\mathcal{L}_M \tag{6}$$

其中, γ是一个可手动调节参数, 用以调和两个模块之间的损失平衡, $\gamma>0$. 与传统的基于特征生成的直推式零

样本学习模型相比, FG-IDA 模型不仅能够通过间接域适应缓解“相互制衡”效应, 还无须根据生成的特征重新训练分类器, 而是直接利用分类网络 C 对样本进行分类. 在测试阶段, FG-IDA 模型可以通过下列操作直接获得给定样本的分类结果:

$$\tilde{y}_i = C(M(x_i)) \tag{7}$$

2.4 模型实施细节

本文遵从常见的基于特征生成的零样本学习方法的特征设置[8,40], 采用了 ResNet101 网络顶层池化层输出的 2 048 维特征作为样本的视觉特征. 对各数据集提取的特征不做任何预处理, 或者数据增强处理.

在特征生成模块部分, 判别器 D 的输入为 d 维特征向量拼接对应 o 维属性向量, 隐藏层只有一层, 是带 LeakyReLU 激活函数的 4 096 维全连接层, 网络输出是一个能够反映样本真假性的 1 维分值, 不做激活. 生成器 G 的输入为 o 维高斯噪声加 o 维属性向量, 后接带 LeakyReLU 激活函数的 4 096 维全连接隐藏层一层, 输出层为对应的 d 维特征向量, 使用 ReLU 作为激活函数. 在间接域适应模块部分, 投影网络输入为 d 维特征向量接一层带 LeakyReLU 激活函数的 4 096 维全连接隐藏层, 输出 q 维嵌入特征, 使用 ReLU 激活. 域分类网络的输入为 q 维嵌入特征, 连接带 ReLU 激活函数的 4 096 维全连接隐藏层, 输出为 2 维 Softmax 标签. 类别分类网络与域分类网络设置相同, 将输出换为未见类类别数. 本文中, q=512, d=2048. 在梯度转置处, M 输出的嵌入特征 h 在进入域分类网络前需要经过一层梯度转置层, 转置层在特征的前向传播时不对特征做任何变动; 但在反向传播时, 则需要乘以系数$-\beta$后再传递给 M. M 输出的嵌入特征进入分类网络则不做任何额外操作.

在参数方面, 特征生成模块分引入的 C-WGAN-GP 部分, 梯度惩罚系数λ设定为 10. 优化器使用 Adam 优化算法, 学习率为 0.000 1. 在特征生成部分, 对抗训练的判别器迭代次数为 5. 在梯度转置处, 为了屏蔽域分类网络在前期未捕捉到分布差异时回传的噪声, 需减小前期的梯度权重. 本文的域分类系数β用下式设定:

$$\beta = \frac{2}{1+\exp(-10\times w)} - 1,$$

其中, w 为训练进度的指示数值.

2.5 “相互制衡”效应

如前文所述, 传统的直推式生成模型通常是一种 FG-DDA 模型, 该模型直接在原特征空间下进行域适应的做法, 往往会引起语义映射和域适应两个任务在优化过程中产生“相互制衡”效应. 这将导致原本在可见类上学习到的语义映射遭受域适应优化的影响甚至破坏. 同样地, 域适应的优化也会受到语义映射学习的影响, 因为本质上语义到特征映射学习的目标是和它相悖, 域适应希望剥离一些具有域信息的细节信息, 而语义映射则会融入一些与域相关的细节信息, 使生成的样本更逼真.

本节将通过设定一些定性实验对“相互制衡”效应进行实验验证. 在这些定性实验中, 本文设置两个比照模型, 如图 3 所示, 第 1 个模型是 FG-IDA 模型的基础生成模型 C-WGAN-GP(以下简称 WGAN), 该模型只有语义映射步骤没有特征域适应步骤; 第 2 个模型本文称为 WGAN-Trans 模型, 该模型将 FG-IDA 模型的投影网络剔除, 直接在视觉空间添加未见类判别器和分类器, 对未见类生成样本进行真实性判别和分类, 即直接在视觉空间实现特征域适应, 因此, WGAN-Trans 模型也可以视为一种与本文提出的 FG-IDA 模型高度相关的 FG-DDA 模型. 为了方便直接从图表中直观地理解实验结果, 后文将 WGAN-Trans 模型称为 FG-DDA 模型. 实验通过分析三者之间生成样本和真实样本的分别差异来分析“相互制衡”效应.

为了验证“相互制衡”效应中在同一空间下特征域适应对语义映射学习(特征生成部分)的影响, 本文将 WGAN, FG-DDA 和 FG-IDA 模型均训练至收敛, 随后固定生成器分别生成数个可见类的生成样本, 并从每个可见类中取出对应的真实样本. 将 WGAN, FG-DDA 与 FG-IDA 这 3 个模型得到的每个类生成样本集分别取中心点, 分别记作 $\hat{x}_{WGAN}$, $\hat{x}_{FG\text{-}IDA}$, $\hat{x}_{FG\text{-}DDA}$; 同时, 为每个可见类的真实样本集同样取中心点, 记作 x_{real}. 将同一个类中的 $\hat{x}_{WGAN}$, $\hat{x}_{FG\text{-}DDA}$, $\hat{x}_{FG\text{-}IDA}$ 分别与 x_{real} 求欧氏距离, 得到每个模型的生成样本中心点与真实样本中心点的欧氏距离, 记作 $dist_{WGAN}$, $dist_{FG\text{-}DDA}$, $dist_{FG\text{-}IDA}$. 该距离一定程度上反映了生成样本与真实样本的分布差异. 本文

通过下列公式定义相对距离差异比值，来定量地分析不同生成模型生成样本与真实样本的分布一致性:

$$\text{相对距离差异比值}=\frac{dist_M-dist_{WGAN}}{dist_{WGAN}},$$

其中，$dist_M$=$dist_{FG\text{-}DDA}$ 或 $dist_M$=$dist_{FG\text{-}IDA}$. 相对距离差异比值越大，说明生成模型生成样本与真实样本分布偏差越大. 如果其值为负，则表示进行相应特征生成模型比 WGAN 模型具有更好真实样本拟合能力.

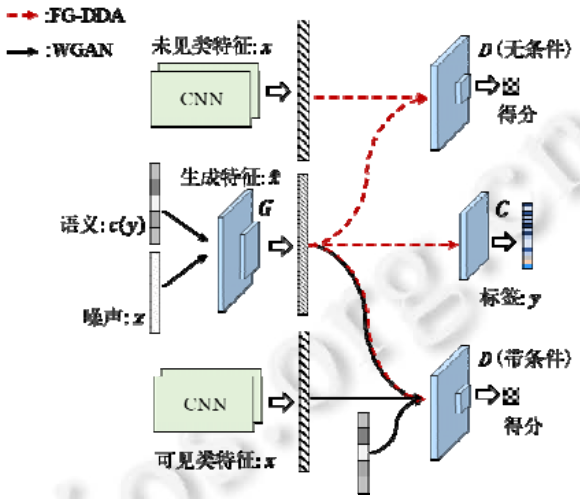


图 3　WGAN 和 FG-DDA 模型结构示意图

本文在 AWA1 数据集上计算了 20 个可见类的 FG-IDA 以及 FG-DDA 模型的相对距离差异比值，以该实验方式验证在同一特征空间下，直接域适应对语义到特征映射建立的负面影响. 实验结果如图 4 所示.

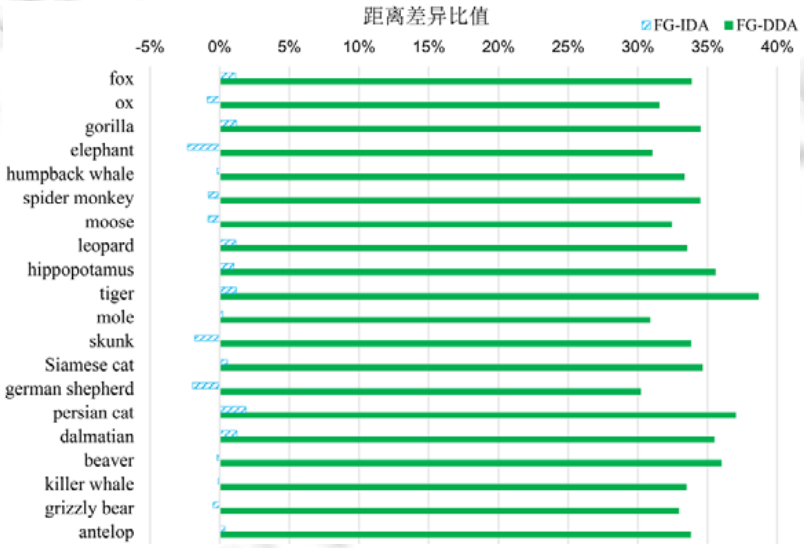


图 4　特征空间下的差异比值图, FG-IDA 和 FG-DDA 与 WGAN 对比

实验结果表明: $dist_{WGAN}$ 和 $dist_{FG\text{-}IDA}$ 并没有太大差别，差异比值仅在–5%–5%之间浮动；而 $dist_{FG\text{-}DDA}$ 则整体比 $dist_{WGAN}$ 增长了 30%–40%. 该现象反映了以下事实: FG-IDA 模型在进行完间接域适应任务以后，在可见类上建立的语义映射并没有受到多少影响，这使得 FG-IDA 生成器得到的生成样本与真实样本的距离与

WGAN 相差无几; 而 FG-DDA 模型在进行了直接域适应之后, 原本在可见类上建立的语义映射遭到了破坏, 使得生成器得到的可见类生成样本与真实的样本距离被拉远. 这证明: 相比于直接域适应, 本文提出的间接域适应方法极大地缓解了域适应任务对可见类语义映射学习过程中的影响.

在“相互制衡”效应中, 除了存在域适应影响语义映射学习过程的现象之外, 同时也存在语义映射也影响域适应的优化过程. 本文以 CUB 数据集为例, 通过可视化生成与真实数据分布的形式, 直观地说明了语义映射对特征域适应的影响. 在实验中, FG-DDA 和 FG-IDA 模型均经过 100 轮次的训练后将模型固定. 首先, 从数据集中取出某个类别的全部真实视觉特征样本; 随后, 使用 FG-DDA 模型生成相同数量的该类别视觉特征; 最后, 将两组特征进行 t-SNE 降维聚类并可视化. 对 FG-IDA 采取相似操作, 不同之处在于, 得到该类别生成特征与真实特征以后, 需分别输入投影网络得到各自的嵌入特征, 然后将生成嵌入特征与真实嵌入特征进行 t-SNE 降维聚类并可视化. 经过上述过程, 最终得到可视化结果, 图 5 是随机挑选的 3 个类别结果(定义行是 FG-IDA 模型在特征空间下的分布图, 底部是 FG-DDA 模型在嵌入空间下的分布图. 圆形为真实特征分布, 方形为生成特征分布).

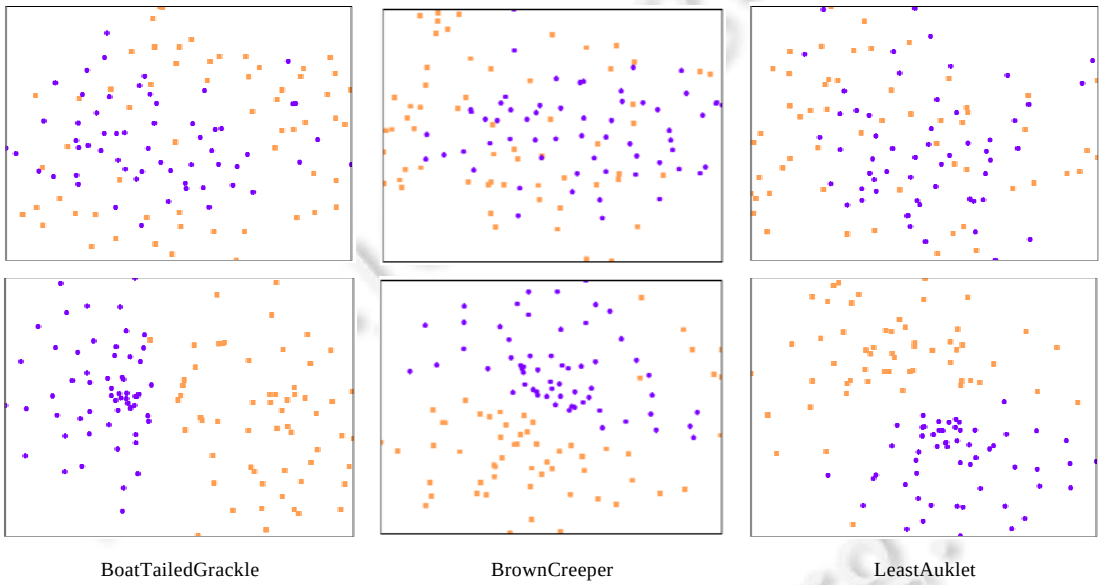
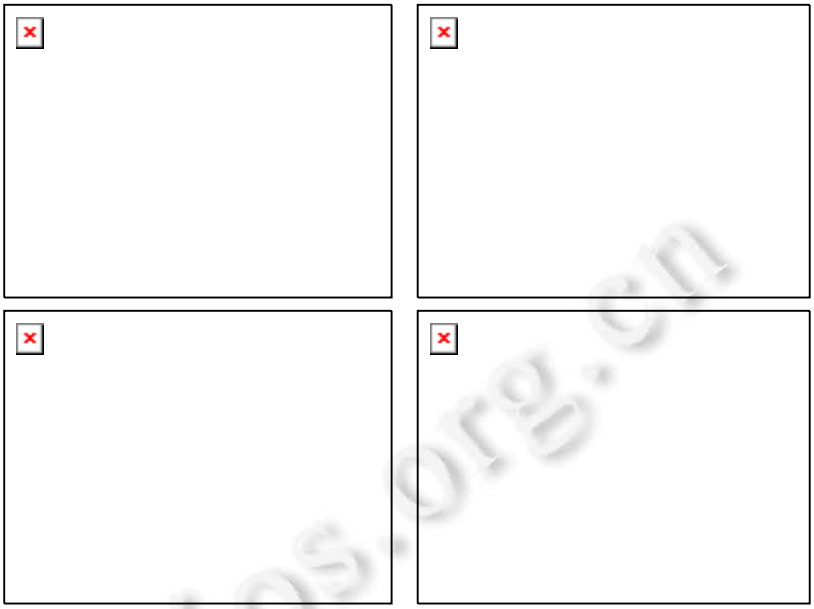


图 5 CUB 数据集上不同类别真实特征与生成特征的 t-SNE 可视化结果

如图 5 所示, 在 CUB 数据集中随机挑选的 3 个类的数据可视化结果中均可发现: FG-DDA 模型的原特征空间下生成的样本和真实样本在分布上有不小的差异, 真实特征与生成特征分别聚集在一起, 两者几乎没有相交; 而在 FG-IDA 模型的嵌入空间下, 真实特征与生成特征虽未完全重合, 但已基本处于同一分布下. 由于生成模型需要输入随机噪声作为“种子”用于特征生成从而引入随机性, 对于同一类别和同一生成模型得到的生成特征会存在一定差异性. 如图 6 所示, 以“LeastAuklet”为例, 使用同一生成模型生成多个批次的特征, 虽然每个批次生成的特征在进行了 t-SNE 降维聚类后, 可视化结果存在一定差异(无论是在 FG-DDA 还是在 FG-IDA 模型上), 但每次实验观测到的特征分布形式仍然与图 5 类似, 即 FG-IDA 模型数据拟合明显优于 FG-DDA 模型. 此外, 除了“LeastAuklet”之外, 在其他类别中, 也可以观测到类似的可视化结果. 这些实验说明: FG- DDA 模型无法很好地完成生成数据与真实数据分布对齐, 而 FG-IDA 模型在嵌入空间下却能取得较好的对齐效果. 这一实验现象也侧面反映了在同一空间下语义映射在学习过程会影响特征域适应的效果, 使得 FG- DDA 模型得到的未见类生成特征与真实特征有较大差异. FG-IDA 模型通过在不同特征空间对语义映射和域适应分别进行优化, 从而简单而又巧妙地调和两个任务优化冲突, 缓解了“相互制衡”效应(顶部是

FG-IDA 模型在特征空间下的分布图, 底部是 FG-DDA 模型在嵌入空间下的分布图. 圆形为真实特征分布, 方形为生成特征分布).



第 1 次实验(生成的第 1 批次特征) 第 2 次实验(生成的第 2 批次特征)

图 6 在 CUB 数据集上, 利用生成特征模型生成两批次不同特征的 t-SNE 可视化结果 (以“LeastAuklet”类别为例)

根据上述实验, 实验结果验证了对于传统基于特征生成的直推式零样本学习方法, 即 FG-DDA 模型, 存在着影响其性能的“相互制衡”效应, 同时也部分验证了 FG-IDA 模型的核心思路——间接域适应, 在面对该效应表现出的先进性.

3 实 验

3.1 数据集与实验设置

本文在 Caltech-UCSD-Birds (CUB)[41], SUN Attribute (SUN)[42]以及 Animals with Attributes 1 (AWA1)[15]和 Animals with Attributes 2 (AWA2)[3]这 4 个常用零样本学习数据集进行实验分析.

CUB, SUN 都是细粒度的数据集: CUB 数据集包含来自 200 种不同鸟类的 11 788 张图像, 每类都具有 312 维的属性描述; SUN 包含来自 717 个场景的 14 340 张图像, 每类具有 102 维属性描述. 而 AWA1 和 AWA2 则是粗粒度数据集, 分别具有 30 475 张图像和 37 322 张图像, 具有 50 个类别, 每个类具有 85 维属性描述.

表 1 给出了各个数据集的详细统计信息. 对上述所有数据集, 均采用约定的标准数据划分方式划定训练与测试集[43].

表 1 各数据集详情(Y_s 为训练集, Y_u 为测试集)

数据集	属性	$Y_s \cup Y_s$	Y_s	Y_u
CUB	312	200	150	50
SUN	102	717	645	72
AWA	85	50	40	10

本文对普通零样本学习方法采用标准实验设置或称为归纳式(inductive)实验设置, 即训练阶段只有可见类所有信息和未见类语义信息可运用. 对于直推式零样本学习方法, 本文采用直推式(transductive)实验设置,

即训练阶段除了可见类所有信息以外, 未见类语义信息和未标定视觉信息也能被运用. 本文提出的 FG-IDA 模型属于直推零样本学习方法, 因此采用直推式实验设置. 与其他零样本学习工作设置[3,22]一样, 在评估方法上,本文采用各类 Top-1 分类准确率 acc_{y^u} 的均值作为零样本学习精度度量:

$$T1=\frac{1}{|Y^u|}\sum acc_{y^u}.$$

本文选定了近年来 15 种经典零样本学习方法作为比对方法, 它们分别为 ALE[2], CLSWGAN[22], LisGAN[44], AREN[45], E-PGN[46], ISE-GAN[47], DSRL[35], GFZSL[23], ALE-Tran[3], QFSL[33], F-VAEGAN-D2[8], SABR-T[40], GMN[48], GXE[49], ADA[13]. 其中, 后 9 种方法属于直推式零样本学习方法.

3.2 实验结果比较

表 2 统计了零样本学习方法在 4 个不同数据集上的精度(Top-1 分类正确率(%)), I 代表 Inductive 设置结果, T 代表 Transductive 设置结果. 粗体数字表示最优结果, 斜体数字表示次优结果.

表 2 FG-IDA 模型与当前的零样本学习方法的性能比较

模型		会议/期刊	AWA1	AWA2	CUB	SUN
I	ALE	TPAMI-2016	59.9	52.5	54.9	58.1
	CLSWGAN	CVPR-2018	68.2	–	57.3	60.8
	LisGAN	CVPR-2019	70.6	–	58.8	61.7
	AREN	CVPR-2019	–	86.7	70.7	61.7
	E-PGN	CVPR-2020	74.4	73.4	72.4	–
	ISE-GAN	WACV-2020	68.4	65.6	63.9	64.7
T	DSRL	CVPR-2017	74.7	72.8	48.7	56.8
	GFZSL	ECML-2017	48.1	78.6	50.0	64.0
	ALE-Tran	CVPR-2017	70.7	–	54.5	55.7
	QFSL	CVPR-2018	–	84.8	69.7	61.7
	F-VAEGAN-D2	CVPR-2019	–	*89.8*	71.1	**70.1**
	SABR-T	CVPR-2019	–	88.9	74.0	67.5
	GMN	CVPR-2019	82.5	–	64.6	64.3
	GXE	CVPR-2019	*89.8*	83.2	61.3	63.5
	ADA	WACV-2020	–	78.6	*74.2*	65.5
	Ours	–	**92.2**	**94.5**	**74.7**	**69.7**

从实验结果可以看出, FG-IDA 模型在 AWA1, AWA2 和 CUB 这 3 个数据集上均取得了当前最优的结果(the state-of-the-art results), 在 SUN 数据集上也取得第二名的好成绩, 且只比最优方法的精度低了 0.4%. 在 AWA1, AWA2 与 CUB 数据集上, FG-IDA 模型分别比相应数据集上排名第二的方法提升了 2.4%, 4.7%, 0.5%的精度. 特别值得一提的是, FG-IDA 模型是唯一在 AWA1, AWA2 这两个数据集上精度均超过 90%的方法. 同时, 作为直推式零样本学习方法, FG-IDA 模型和近年来提出的其他直推式方法相比优势比较明显. 如 ADA 是 2020 年提出的最新直推式零样本学习方法之一, FG-IDA 模型在 AWA2, CUB, SUN 数据集上分别比其取得 15.9%, 0.5%, 4.2%的精度提升. 尤其是与作为典型直接域适应的 F-VAEGAN-D2 模型相比, FG-IDA 模型在 AWA2 和 CUB 数据集上提升了 4.7%和 3.6%, 在 SUN 上仅相差 0.4%. 上述实验现象验证了 FG-IDA 模型的有效性, 同时也验证了间接域适应思想的先进性.

另一个有趣的实验现象, 直推式零样本学习方法在性能往往优于归纳式(标准)零样本学习方法. 比如同为 2019 年 CVPR 会议录用的工作, 直推式方法 SABR-T 在 AWA2, CUB, SUN 这 3 个数据集上分别比归纳式方法 AREN 分别提高了 2.2%, 3.3%, 5.8%的精度. 本文认为: 这是由于在直推式实验设置中, 未见类的数据的应用能够一定程度上缓解映射域漂移问题, 从而具备更好的性能.

实验结果还表明, 生成模型在零样本学习中具备更优异性能. 如在归纳式设置, 生成模型 E-PGN 在 AWA1, AWA2, CUB 数据集上均取得最优的结果. 同时, 在直推式设置中, 除了本文提出的 FG-IDA 模型以外, 表现最为优异的零样本学习方法 F-VAEGAN-D2 也是一种基于特征生成的方法. 本文把生成模型的优异表现归结于其优秀数据分布拟合能力.

从表 2 中我们还可以看到, 有一些非直推式的方法效果要好于直推式的方法, 且除了 GFZSL 方法外, 该情况均集中出现在 CUB 和 SUN 数据集上, 其中以 CUB 数据集最为突出. 这是因为 AWA 数据集为粗粒度数据集, 而 CUB 和 SUN 为细粒度数据集, 区别相对较小, 个体差别均体现在细节上, 并且细粒度的数据集属性向量更高, 语义信息更丰富. 因此, 一些对属性开发特别重视的方法如 E-PGN, AREN(引入了注意力机制), 会在这些数据集上具有更大的优势. 而本文提出的 FG-IDA 模型虽然未使用注意力机制, 但取得的效果仍然高于前两种方法, 这也在一定程度上体现了间接域适应的有效性.

除了以上所述之外, 本文提出的 FG-IDA 还具有最为良好的推广能力. 无论是在归纳式方法还是直推式方法中, 许多方法均存在部分数据集过拟合现象严重的问题. 例如 ADA 模型, 其在 CUB 数据集上取得了次优精度, 但在 AWA2 数据集上却甚至低于一些归纳式方法. 而 FG-IDA 模型能在 3 个数据集上取得最优结果, 在 SUN 上也能取得不俗的次优结果.

3.3 消融实验

本节将在 CUB 和 SUN 两个数据集上对 FG-IDA 模型进行消融分析(ablation analysis), 以探讨间接域适应与端到端学习模式的有效性. 在消融实验中, 本节设置 3 个对比组, 分别为 WGAN, FG-DDA 以及 FG-IDA 非端到端版本或两阶段分别优化版本(two-stage). 其中, WGAN 模型是 FG-IDA 模型除去域适应模块模型的基础特征生成模型 C-WGAN-GP, 它可被视为一个归纳式方法. FG-DDA 模型和 FG-IDA 模型在使用同样的生成模块, 只是前者移除了投影网络, 使得类别分类和域分类(判别器)直接在视觉特征空间进行.

如第 2.5 节所述, FG-DDA 可被视为 FG-IDA 模型的直接域适应版本. FG-IDA (two-stage)模型指的是特征生成和域适应两个模块独立优化, FG-IDA(end-to-end)模型指特征生成模块与域适应模块以合适权重桥接,并共同优化. 表 3 记录了 FG-IDA 模型与其对比组的零样本学习的性能.

表 3 各生成模型零样本学习精度比对(%)

数据集	WGAN	FG-DDA	FG-IDA (two-stage)	FG-IDA (end-to-end)
CUB	54.4	69.9	70.1	**74.5**
SUN	52.2	65.7	65.9	**69.7**

表 3 中的结果表明, 无论是 FG-DDA 还是 FG-IDA 模型, 其表现均明显优于 WGAN. 如 FG-DDA 模型在 CUB 和 SUN 数据集上分别比 WGAN 多取得 15.5%和 13.5%的零样本分类精度. FG-IDA 模型在这些数据集上的性能提升更为明显, FG-IDA(end-to-end)模型在两个数据集上均提升超过 20%的精度. 这些现象说明了直推式方法可以利用未见类未标定数据缓解映射域漂移问题, 从而提升零样本学习的性能.

FG-IDA 模型在 CUB 和 SUN 数据集上均表现比 FG-DDA 模型更为优异, 它相对 FG-DDA 模型在 CUB 和 SUN 数据集上分别提升 4.6%和 4.0%分类精度. 在前文中所述, FG-DDA 和 FG-IDA 模型的差异是 FG-IDA 模型引入了投影网络以建立一个嵌入空间进行未见类生成特征的间接域适应, 而 FG-DDA 则是直接在原特征空间进行未见类生成特征的域适应. 因此, 上述实验结果证明了间接域适应相较于直接域适应在零样本学习上的优越性和先进性.

本节也对 FG-IDA 模型进行端到端学习还是独立学习进行了讨论. 表 3 中的实验结果表明, FG-IDA (end-to-end)比 FG-IDA (two-stage)分别在 CUB 和 SUN 数据集上提升了 4.4%和 3.8%的精度. 上述现象说明: 端对端学习模式能够从全局角度更好地优化 FG-IDA 模型各个步骤, 进而充分发挥其性能.

除了上述实验现象外, 表 3 记录的结果中还反映了一个有趣的现象, 即 FG-DDA 与 FG-IDA (two-stage)模型取得了相似的性能, 甚至 FG-IDA (two-stage)模型在性能上比 FG-DDA 模型在精度上有略微优势. 这说明了 FG-DDA 模型的端到端学习模式并未对语义映射与域适应进行良好的优化, 甚至因为“相互制衡”效应, FG-DDA 模型统一优化的效果还不如对语义映射与域适应部分单独进行优化的效果好. 如, FG-IDA (two-stage)模型比 FG-DDA 模型在两个数据集上的零样本学习精度均提升 0.2%.

3.4 “相互制衡”效应实验分析

在第 2.5 节实验的基础上，为了进一步验证“相互制衡”效应，本文扩充了第 3.2 节的实验，在 CUB 和 SUN 数据集上进一步实验比对了 WGAN, FG-DDA 和 FG-IDA 模型在未见类和可见类数据上的分类精度. 表 4 记录了这些实验结果.

表 4　各模型在可见类与未见类数据的分类精度比对(Top-1 精度(%))

模型	CUB		SUN	
	可见类	未见类	可见类	未见类
WGAN	**79.2**	53.1	**55.1**	49.1
FG-DDA	67.3	69.9	31.9	65.7
FG-IDA	78.9	**74.5**	54.4	**69.7**

如表 4 所示，通过比对 WGAN 和 FG-DDA 模型的实验结果，可以发现 FG-DDA 模型恶化了 WGAN 模型在可见类数据的分类精度. 在可见类数据上，FG-DDA 模型在 CUB 和 SUN 数据集上分别比 WGAN 模型低了 11.9%和 23.2%的分类精度；而 FG-IDA 模型和 WGAN 模型在可见类数据上表现几乎持平，在 CUB 和 SUN 数据集上仅仅分别低了 0.3%和 0.7%的分类精度. 这一实验现象说明了，FG-DDA 模型的直接域适应模式相较于 FG-IDA 模型的间接域适应模式更容易影响语义到视觉特征映射的建立(特征生成模块的优化)，即直接域适应对语义映射的优化存在着制衡效应. 而 FG-IDA 的间接域适应能够很好地缓解这一问题，从而最大限度地避免对映射建立优化任务的影响.

此外，表 4 中的未见类数据上的实验结果还揭示了一个重要的现象，即 FG-IDA 模型相较于 FG-DDA 模型对 WGAN 模型上分类精度提升更大. 如 FG-IDA 模型相对 FG-DDA 模型在 CUB 和 SUN 数据集上分别多提升 4.6%和 4.0%零样本分类精度(未见类数据分类精度). 这一现象说明了 FG-DDA 模型并未最优化特征域适应模块. 这是因为在同一空间中映射建立的优化可能对域适应存在着反向制衡效应，导致 FG-DDA 无法发挥特征域适应模块的最佳效果. 同样地，FG-IDA 模型通过引入一个嵌入空间独立优化特征域适应，能够很好地调和上述制衡效应，从而进一步提升性能.

3.5 γ参数实验分析

γ是 FG-IDA 模型中唯一手动调节的参数，用于调和特征生成模块与间接域适应模块的损失. 图 7 展示了不同γ值的设置在不同数据集上对 FG-IDA 模型的性能影响.

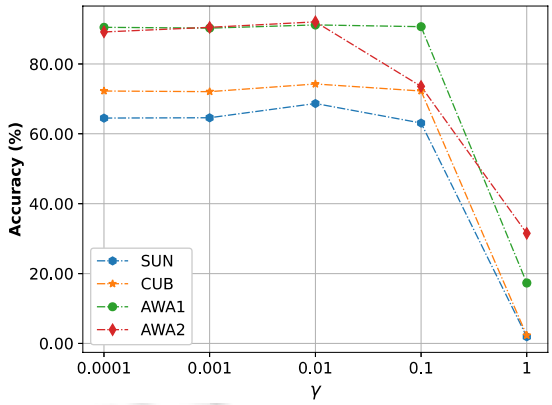


图 7　γ的取值在各个数据集上对 FG-IDA 模型的零样本学习性能的影响示意图

如图 7 所示，实验结果表明，当γ的取值在 10^{-4} 到 10^{-1} 之间, FG-IDA 模型表现平稳且均能取得较为优异的分类精度；当γ取值超过 10^{-1} 后, FG-IDA 模型的性能则急剧下降. 这一实验现象说明，当过度强调特征域适应的优化，必然会影响特征生成步骤的优化，从而导致模型奔溃. 从模型角度对上述现象也可以给出合理的解

释: FG-IDA 模型作为一种利用数据增强解决零样本学习的方法, 其重点仍然是在特征生成模块, 即语义映射的建立, 而间接域适应模块只是辅助其更好地生成具有更好推广能力和判别能力的特征用于零样本学习. 因此, γ作为平衡间接域适应模块和特征生成模块的重要参数, 不宜被设置过大.

图 7 的实验结果还揭示了 FG-IDA 模型在参数设置的另一个优点: FG-IDA 模型最优γ参数值对数据集选择不敏感. FG-IDA 模型在 4 个数据集上γ的最优参数值均为 10^{-2}. 因此, FG-IDA 模型在各数据集上的参数γ可设定为固定值.

4 总 结

本文通过对传统基于特征生成的直推式零样本学习方法分步骤实验分析, 发现了在同一特征空间下进行语义映射与特征域适应两项优化任务的做法存在着“相互制衡”效应, 从而导致模型无法发挥最佳的零样本学习性能. 为了解决上述问题, 本文提出了间接域适应特征生成(feature generation with indirect domain adaption, FG-IDA)模型. 该模型通过为特征域适应模块引入一个嵌入空间, 串行化语义映射与特征域适应的优化步骤, 从而保证两个任务在不同空间进行优化, 缓解两者之间的“相互制衡”效应. 本文在 AWA1, AWA2, CUB, SUN 这 4 个经典零样本数据集上的大量实验结果, 验证了“相互制衡”效应的存在性以及 FG-IDA 模型的有效性与先进性. 特别是在 AWA1, AWA2, CUB 这 3 个数据集上, FG-IDA 模型取得了当前最优(the state-of-the-art)的零样本学习精度.

References:

[1] Lampert CH, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. In: Proc. of the CVPR. 2009. 951−958.

[2] Akata Z, Reed S, Walter D, *et al*. Evaluation of output embeddings for fine-grained image classification. In: Proc. of the CVPR. 2015. 2927−2936.

[3] Xian Y, Schiele B, Akata Z. Zero-shot learning—The good, the bad and the ugly. In: Proc. of the CVPR. 2017. 3077−3086.

[4] Lampert CH, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. Pattern Analysis and Machine Intelligence, 2014, 36(3): 453−465.

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

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

[7] Fu Z, Xiang T, Kodirov E, *et al*. Zero-shot object recognition by semantic manifold distance. In: Proc. of the CVPR. 2015. 2635−2644.

[8] Xian YQ, Sharma S, Schiele B, *et al*. f-VAEGAN-D2: A feature generating framework for any-shot learning. In: Proc. of the CVPR. 2019. 10275−10284.

[9] Schonfeld E, Ebrahimi S, Sinha S, *et al*. Generalized zero- and few-shot learning via aligned variational autoencoders. In: Proc. of the CVPR. 2019. 8247−8255.

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

[11] Yang L, Jing LP, Yu J. Heterogeneous transductive transfer learning algorithm. Ruan Jian Xue Bao/Journal of Software, 2015, 26(11): 2762−2780 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4892.html [doi: 10.13328/j.cnki.jos.004892]

[12] Ji Z, Sun T, Yu YL. Transductive discriminative dictionary learning approach for zero-shot classification. Ruan Jian Xue Bao/Journal of Software, 2017, 28(11): 2961−2970 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5338.html [doi: 10.13328/j.cnki.jos.005338]

[13] Khare V, Mahajan D, Bharadhwaj H, *et al*. A generative framework for zero shot learning with adversarial domain adaptation. In: Proc. of the WACV. 2020. 3101−3110.

[14] Wu J, Zhang T, Zha ZJ, *et al*. Self-supervised domain-aware generative network for generalized zero-shot learning. In: Proc. of the CVPR. 2020. 12767−12776.

[15] Akata Z, Perronnin F, Harchaoui Z, *et al*. Label-embedding for attribute-based classification. In: Proc. of the CVPR. 2013. 819−826.

[16] Al-Halah Z, Tapaswi M, Stiefelhagen R. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In: Proc. of the CVPR. 2016. 5975−5984.

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

[18] Kankuekul P, Kawewong A, Tangruamsub S, *et al*. Online incremental attribute-based zero-shot learning. In: Proc. of the CVPR. 2012. 3657−3664.

[19] Yang G, Liu JL, Li XR, *et al*. Visual feature combination approach for zero-shot learning. Ruan Jian Xue Bao/Journal of Software, 2018, 29: 16−29 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18014.htm

[20] 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.

[21] Chen L, Zhang H, Xiao J, *et al*. Zeroshot visual recognition using semantics-preserving adversarial embedding networks. arXiv: 1712.01928, 2017.

[22] Xian Y, Lorenz T, Schiele B, *et al*. Feature generating networks for zero-shot learning. In: Proc. of the CVPR. 2018. 5542−5551.

[23] Verma VK, Rai P. A simple exponential family framework for zero-shot learning. In: Proc. of the ECML PKDD. Cham: Springer, 2017. 792−808.

[24] Wang WL, Pu YC, Verma VK, *et al*. Zero-shot learning via class-conditioned deep generative models. arXiv:1711.05820, 2017.

[25] Li Y, Swersky K, Zemel R. Generative moment matching networks. In: Proc. of the ICML. 2015. 1718−1727.

[26] Gretton A, Borgwardt KM, Rasch MJ, *et al*. A kernel method for the two-sample problem. arXiv:0805.2368, 2008.

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

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

[29] Verma VK, Arora G, Mishra A, *et al*. Generalized zero-shot learning via synthesized examples. In: Proc. of the CVPR. 2018. 4281−4289.

[30] Mishra A, Reddy SK, Mittal A, *et al*. A generative model for zero shot learning using conditional variational autoencoders. In: Proc. of the CVPR. 2018. 2188−2196.

[31] Bucher M, Herbin S, Jurie F. Generating visual representations for zero-shot classification. In: Proc. of the CVPR. 2017. 2666−2673.

[32] Lu J, Li J, Yan ZA, *et al*. Zero-shot learning by generating pseudo feature representations. arXiv:1703.06389, 2017.

[33] Song J, Shen CC, Yang YZ, *et al*. Transductive unbiased embedding for zero-shot learning. In: Proc. of the CVPR. 2018. 1024−1033.

[34] Kodirov E, Xiang T, Fu ZY, *et al*. Unsupervised domain adaptation for zero-shot learning. In: Proc. of the ICCV. 2015. 2452−2460.

[35] Ye M, Guo YH. Zero-shot classification with discriminative semantic representation learning. In: Proc. of the CVPR. 2017. 7140−7148.

[36] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: Proc. of the ICML. 2015. 1180−1189.

[37] Zhang ZM, Saligrama V. Learning joint feature adaptation for zero-shot recognition. arXiv:1611.07593, 2016.

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

[39] Gulrajani I, Ahmed F, Arjovsky M, *et al*. Improved training of Wasserstein GANs. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 5767−5777.

[40] Paul A, Krishnan NC, Munjal P. Semantically aligned bias reducing zero shot learning. In: Proc. of the CVPR. 2019. 7056−7065.

[41] Wah C, Branson S, Welinder P, *et al*. The Caltech-UCSD birds-200-2011 dataset. Technical Report, CNS-TR-2011-001, 2011. http://www.vision.caltech.edu/datasets/cub_200_2011/

[42] Patterson G, Hays J. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proc. of the CVPR. 2012. 2751−2758.

[43] Reed S, Akata Z, Lee H, *et al*. Learning deep representations of fine-grained visual descriptions. In: Proc. of the CVPR. 2016. 49−58.

[44] Li JJ, Jin MM, Lu K, *et al*. Leveraging the invariant side of generative zero-shot learning. arXiv:1904.04092, 2019.

[45] Xie GS, Liu L, Jin XB, *et al*. Attentive region embedding network for zero-shot learning. In: Proc. of the CVPR. 2019. 9384−9393.

[46] Yu YL, Ji Z, Han JG, *et al*. Episode-based prototype generating network for zero-shot learning. In: Proc. of the CVPR. 2020. 14035−14044

[47] Pambala AK, Dutta T, Biswas S. Generative model with semantic embedding and integrated classifier for generalized zero- shot learning. In: Proc. of the WACV. 2020. 1237−1246.

[48] Sariyildiz MB, Cinbis RG. Gradient matching generative networks for zero-shot learning. In: Proc. of the CVPR. 2019. 2168−2178.

[49] Li K, Min MR, Fu Y. Rethinking zero-shot learning: A conditional visual classification perspective. In: Proc. of the CVPR. 2019. 3583−3592.

附中文参考文献:

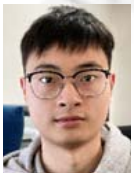
[11] 杨柳, 景丽萍, 于剑. 一种异构直推式迁移学习算法. 软件学报, 2015, 26(11): 2762−2780. http://www.jos.org.cn/1000-9825/4892.html [doi: 10.13328/j.cnki.jos.004892]

[12] 冀中, 孙涛, 于云龙. 一种基于直推判别字典学习的零样本分类方法. 软件学报, 2017, 28(11): 2961−2970. http://www.jos.org.cn/1000-9825/5338.html [doi: 10.13328/j.cnki.jos.005338]

[19] 杨刚, 刘金露, 李锡荣, 许洁萍. 一种基于视觉特征组合构造的零样本学习方法. 软件学报, 2018, 29: 16−29. http://www.jos.org.cn/1000-9825/18014.htm

黄晟(1988－), 男, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为模式识别, 机器学习, 计算机视觉.

张小洪(1973－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为图像处理, 机器学习, 模式识别, 数据挖掘, 智能软件工程.



杨万里(1996－), 男, 硕士生, 主要研究领域为计算机视觉.

杨丹(1962－), 男, 博士, 教授, 博士生导师, 主要研究领域为模式识别, 图像处理, 软件工程, 智能制造.

张译(1994－), 男, 博士生, 主要研究领域为计算机视觉, 深度学习.