基于类别扩展的广义零样本图像分类方法

张 杰[1],廖盛斌[2],张浩峰[1],陈得宝[3]
(1. 南京理工大学计算机科学与工程学院,江苏南京 210094; 2. 华中师范大学国家数字化学习工程技术研究中心,湖北武汉 430079; 3. 淮北师范大学计算机科学与技术学院,安徽淮北 235000)

摘 要: 在传统的零样本图像分类方法中,语义属性通常被用作辅助信息来描述各类别的视觉特征. 然而,单一的语义属性并不能对类内多样性的视觉特征进行全面的描述. 为提高语义属性对类别内部多样性的表示能力,同时也为了帮助模型提高对各类别的描述能力,本文通过属性自编码器的方式在视觉以及语义空间上对类别进行扩展. 此外,为了缓解传统生成性方法因无法直接计算生成空间到真实空间的变换而带来的模型次优解问题,本文采用了生成流网络作为基础网络,通过可逆变换直接计算两个空间之间的变换来开展对零样本学习任务的研究. 本文使用解码器网络将逆生成流网络为测试样本生成的原型特征解耦成视觉原型及语义原型信息,然后根据这两个原型信息实现将测试样本预分类到可见类集或不可见类集中,最终在这两个子分类空间中根据样本的特点分别进行监督分类和零样本分类任务以提高模型的整体性能表现. 本文在五个数据集上通过大量的实验验证了本文所提方法的有效性.

关键词: 生成流; 预分类; 广义零样本学习; 类扩展
基金项目: 国家自然科学基金(No.61872187,No.62077023,No.62072246);江苏省自然科学基金(No.BK20201306)
中图分类号: TP391.4;TP37 **文献标识码:** A **文章编号:** 0372-2112(2023)04-1068-13
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220036

Category Expansion Based Generalized Zero-Shot Image Classification

ZHANG Jie[1], LIAO Sheng-bin[2], ZHANG Hao-feng[1], CHEN De-bao[3]
(1. *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu* 210094, *China*;
2. *National Digital Learning Engineering Technology Research Center, Central China Normal University, Wuhan, Hubei* 430079, *China*;
3. *School of Computer Science and Technology, Huaibei Normal University, Huaibei, Anhui* 235000, *China*)

Abstract: In traditional zero-shot image classification methods, semantic attributes are usually used as auxiliary information to describe the visual features of each class. However, a single semantic attribute cannot fully describe the diverse visual features within a single class. To improve the ability of semantic attributes to express the diversity within the class, and to help the model improve the description ability for each category, we use the semantic auto-encoder to expand the categories in visual and semantic space. In addition, to alleviate the suboptimal solution problem of the model caused by the inability to directly calculate the transformation from the generation space to the real space by the traditional generative methods, we employ the generative flow as the basic network in this paper to directly calculate the transformation between the two spaces. Furthermore, we exploit the decoder network to decouple the prototype features generated by the inverse generative flow network for the test samples into visual prototypes and semantic prototypes, and then realize the pre-classification of the test samples into seen or unseen classes. Finally, in the two sub-classification domains, supervised classification and zero-shot classification are performed separately to improve the overall performance. Extensive experiments are conducted on five popular datasets to verify the effectiveness of the proposed method.

Key words: generative flow; pre-classification; generalized zero-shot learning; category expansion
Foundation Item(s): National Natural Science Foundation of China (No.61872187, No.62077023, No.62072246); Natural Science Foundation of Jiangsu Province (No.BK20201306)

1 引言

作为计算机视觉重点任务之一,针对图像识别与分类的研究也从最初的基于机器学习的传统方法发展到基于深度学习的方法. 随着软硬件技术的不断升级与发展,在有充足标记数据下训练出来的监督分类模型已经能够媲美人类自身对图像的分类识别能力. 但

收稿日期:2022-01-03;修回日期:2022-02-28;责任编辑:朱梅玉

是对大规模数据集进行监督分类仍存在两个问题. 首先,对样本尤其是对那些仅有微小差异的细粒度样本进行大规模标记是一件成本巨大且对标注人员的专业知识有着很强要求的事情. 其次,训练得到的监督模型无法对新类别进行有效的识别. 为了解决传统监督分类模型对缺乏标记样本或新生类别无法进行有效分类的问题,Larochelle 等[1]提出了一种面向无标记样本的无监督分类任务——零样本学习(Zero-Shot Learning, ZSL). 在该任务的设置下,训练集中不会包含任何测试集中所存在类别的样本. 之后,研究者们逐渐注意到ZSL任务设置中所存在的问题与不足,即测试集中也可能会出现训练集中的部分类别的样本并由此推广到了由 Chao 等[2]提出的更加符合现实场景的广义零样本学习(Generalized Zero-Shot Learning,GZSL)任务中去.

为了能对未知的测试集进行识别与分类,研究者们以各类别的语义属性信息作为辅助信息陆续提出了大量的非生成性方法. 例如 Lampert 等[3]从贝叶斯理论出发,提出了经典的直接属性预测(Direct Attribute Prediction,DAP)与非直接属性预测(Indirect Attribute Prediction,IAP)方法,程玉虎等[4]在 DAP 方法的基础上采用混合属性的方式提出了混合属性 DAP 方法. 这些方法虽然能够在传统的零样本学习任务中取得不错的分类效果,但是当应用到广义零样本学习任务中时却有着非常差的性能表现. 例如 IAP 模型的在 AWA1 数据集上的分类准确率从 ZSL 任务中的 35.9% 降低到了GZSL任务中的4.1%. 造成这一现象的根本原因在于模型承受了较大的偏置问题,即在可见类(seen classes)上训练的模型会更加倾向于将不可见类(unseen classes)分类为可见类. 为缓解偏置问题,赵鹏等[5]将可见类以及不可见类在语义空间上的关系迁移到了视觉空间上来保证模型在训练过程中能够兼顾不可见类的信息,Bai 等[6]选择将语义属性和视觉特征通过对偶判别性自编码器映射到一个隐空间中,Zhang 等[7]则通过推导式的训练模式将无标签的不可见类样本也送入模型的训练过程以兼顾不可见类别的相关信息. 而一系列基于生成性方法的研究则从为不可见类合成高质量的伪样本出发来将缺失不可见类样本的零样本分类学习问题转化为传统的监督分类学习问题[8].

我们观察到在样本的视觉特征空间中存在一些类别有着较大类内差异却共享着相同的用于描述视觉特征的语义属性变量的现象. 为了帮助模型更好地理解各类别内的差异,同时也为了缓解语义属性无法描述类别内部存在的视觉差异的问题,本文对原始数据集进行了类别扩展. 首先在视觉空间中采用聚类的方式对各类别进行再分类,通过将原始的单一类别划分为若干新类别来减轻类内差异较大的问题. 接着,为了使得语义属性向量能够对每一个新的类别进行描述,本文在原始的语义属性向量上添加对应数量的可学习的语义噪声数据以构成新的语义属性向量并利用属性自编码器[9](Semantic Auto-Encoder, SAE)的方法来学习这些语义噪声. 最终使得对于数据集中的每一个类别来说,都有着充足的语义属性来描述它的视觉特征.

广泛应用在零样本分类学习中的两种生成性网络——对抗生成网络[10](Generative Adversarial Network, GAN)与变分编码器网络[11](Variational Auto-Encoder, VAE)都有着不错的表现. 然而因为对生成样本空间与真实样本空间分布的一致性度量是一个计算量很大甚至不可解的问题,GAN 与 VAE 这两种生成性模型分别利用 Jensen-Shannon 散度来衡量生成样本与真实样本之间的相似性,以及对一致性度量的下确界函数进行近似求解的方式来避开对分布变换的直接计算. 但显而易见的是,无论是 GAN 还是 VAE 的计算方式都是一种次优解或者说近似解,而这也会对模型的性能带来一定的影响. 与这两种方式不同的是,生成流网络模型[12](Flow Net)利用可逆函数构造激活函数并通过解耦技巧简化对模型的计算从而使得精确计算生成样本空间到真实样本空间的变换成为了可能. Chen 等[13]与 Shen 等[14]也注意到了传统生成性方法的次优性并先后将生成流网络引入到零样本分类学习任务中来.

本文也使用生成流网络作为基础网络来更好地为不可见类生成视觉原型与语义属性原型信息. 并且由于可见类与不可见类样本之间存在一定的固有差异,因此当不可见类集中的样本进入生成流模型中后,它所生成的视觉原型与语义属性原型必然会与由可见类集中的样本所生成的视觉原型及语义原型信息存在着一定的差异. 为了利用这种差异性,本文设计了一个预分类模块来判断生成流模块所接收的样本是否来自不可见类从而将测试样本率先划分到可见类集或不可见类集合中,最终再根据可见类与不可见类样本的特点,在两个子空间中进行最终的分类任务.

本文所提方法主要有如下几个创新点.

(1)为避免模型的次优解问题,本文以生成流网络为基础提出了一种新的零样本分类学习方法来同时学习测试样本的视觉原型信息与语义属性原型信息,并采用一个预分类器模块以这两种原型信息为输入来判断测试样本的归属集.

(2)我们注意到在各类别上广泛存在着用单一的语义属性向量来描述具有较大类内差异的视觉特征的现象. 为缓解这一问题,本文在原始的视觉特征空间上对类别进行了扩充,同时利用自编码器的方法来学习新的语义属性向量.

(3)本文在四个通用数据集上进行了大量的实验,

实验结果表明本文方法在所有数据集上都能取得很好的分类性能.

2 相关工作

随着近些年来对零样本学习任务研究的不断深入,零样本学习在图像分类[15~17]领域受到越来越多的关注. 本节将围绕零样本图像分类学习任务介绍一些近些年的相关工作.

2.1 生成性方法

为解决零样本学习中缺少不可见类别的视觉特征的问题,研究者们基于生成性模型提出了大量方法来为不可见类别生成高质量的伪样本,进而利用这些伪样本与真实的可见类样本一起训练一个监督分类模型,最终使得原始的零样本学习问题成功转变为传统的监督分类学习任务. 在零样本学习中,主流的生成性方法主要基于生成对抗网络和变分编码器器网络,最近几年基于生成流网络的生成性方法也逐渐吸引到了研究者们的注意.

2.1.1 基于对抗网络的生成性方法

Xian等[18]开创性地将传统的生成对抗网络引入到零样本图像分类学习任务中来. 而为了使得模型能够生成质量更高的、更具辨别性的伪样本,他们在GAN网络的判别器上增加了一个分类器分支. Sariyildiz等[19]从提高分类器性能表现的角度出发,通过梯度匹配的方法强制约束生成器所产生的伪样本的质量与分类器的性能高度相关,以保证所产生的伪样本能够使得分类器有着更高的分类表现. Schönfeld等[20]所提出的CADA-VAE方法则通过对齐视觉特征和语义信息的方式学习到一个包含多模态信息的隐空间,并在隐空间中完成对样本的分类任务. 而Wang等[21]将每一个类别表示为一个受到语义属性约束的隐空间,并将这个隐空间的分布信息作为一个先验知识输入到变分编码器中学习一个具有高度辨别能力的特征表示.

2.1.2 基于生成流的方法

Shen等[14]提出了一个名为可逆零样本流(Invertible Zero-shot Flow,IZF)的模型,通过将类别语义作为一种条件约束并利用最大平均差异[22,23](Maximum Mean Discrepancy,MMD)约束调整样本间偏置问题来提高模型的性能表现. 而Chen等[13]则采用一系列条件仿射耦合层组成条件生成流,特别是将扰动注入原始视觉特征以补充潜在模式,并通过相对于语义锚点的相对定位来计算全局语义.

2.2 基于预分类的方法

随着异常点检测方法[24]的日益成熟,一些基于异常点检测的预分类方法在零样本任务中也取得了很好的成绩. 这些方法从样本本身出发,出于可见类与不可见类样本之间存在固有差异的事实,在测试阶段首先对样本进行预分类任务,之后根据各集合中样本的特点分别进行监督分类或者传统零样本分类任务. 而由于这两个待分类空间互为补子空间,基于预分类的方法在一定程度上也缓解了模型的偏置问题.

其中Chen等[25]将样本分别从原始的视觉特征空间与语义属性空间中通过SVAE[26]模型映射到一个超分球空间中去,并在这个隐空间中完成对测试样本的预分类操作,之后分别对分类到可见类集以及不可见类集的样本进行分类. Atzmon等[27]则通过利用软阈值函数来实现对测试样本的二分类任务,同时通过在各模型之间共享信息来提高模型的性能. Min等[28]则通过构建互补的语义无关及语义相关视觉特征,并设计了一个熵检测器来判断测试样本是否来自可见类集.

3 方法

本节对本文所提出的基于类别扩展的零样本学习方法(Category Expansion Based Generalized Zero-Shot Learning,CEBGZSL)的理论及其优化细节做详细的阐述. 而为了能够便于理解,在正式介绍方法模型之前,先将对泛化零样本分类学习的任务设置以及将涉及的相关变量的定义及其含义进行解释说明. 同时也在图1中形象化的展示了CEBGZSL方法. 如图1所示,在类别扩展阶段,本文以“吉娃娃”类为例展示了在视觉空间$\boldsymbol{X}$上根据视觉差异将吉娃娃类扩展为三种类别并通过属性自编码器在语义空间$\overline{\boldsymbol{A}}$上学习到多样性语义属性的过程. 在训练阶段中,本文通过编码器En将各类别的视觉原型以及语义原型映射为一个联合原型,并利用正向生成流网络f将该联合原型变换到由视觉特征经过高斯混合模型GMM得到的视觉空间上. 同时,一个解码器De又将联合原型解耦回样本的视觉原型$\tilde{\boldsymbol{P}}$及语义原型$\tilde{\boldsymbol{A}}$. 而在分类阶段中,测试样本经过逆生成流网络f^{-1}产生伪联合原型,并进一步通过De产生该测试样本的合成视觉原型$\tilde{\boldsymbol{P}}_{\mathrm{u}}$及合成语义原型$\tilde{\boldsymbol{A}}_{\mathrm{u}}$,接着利用这两个原型计算出相应的得分并产生加权得分$\boldsymbol{V}$,最终根据阈值t的大小决定使用哪种分类器为该样本进行分类.

3.1 问题描述与变量定义

设数据集$\mathcal{T}$是一个含有$C+U$个不同类别的数据集. 其子集$\mathcal{T}_s$由C种可见类样本所构成,子集$\mathcal{T}_u$是由U种不可见类样本所构成,且两个子集互补,即$\mathcal{T}_s\cup\mathcal{T}_u=\mathcal{T}$及$\mathcal{T}_s\cap\mathcal{T}_u=\varnothing$. 对可见类集合$\mathcal{T}_s=\left\{\left(\boldsymbol{x}_i,y_i,\boldsymbol{a}_{y_i}\right)|\boldsymbol{x}_i\in\boldsymbol{X}^{n_s\times d},y_i\in\{1,2,\cdots,C\},\boldsymbol{a}_{y_i}\in\boldsymbol{A}^{C\times k}\right\}$,它包含了$C$种类别的$n_s$个样本. 集合中的第$i$个样本的$\boldsymbol{x}_i$表示该样本的一个$d$维视觉特征向量. 设$y_i$为第$i$个样本

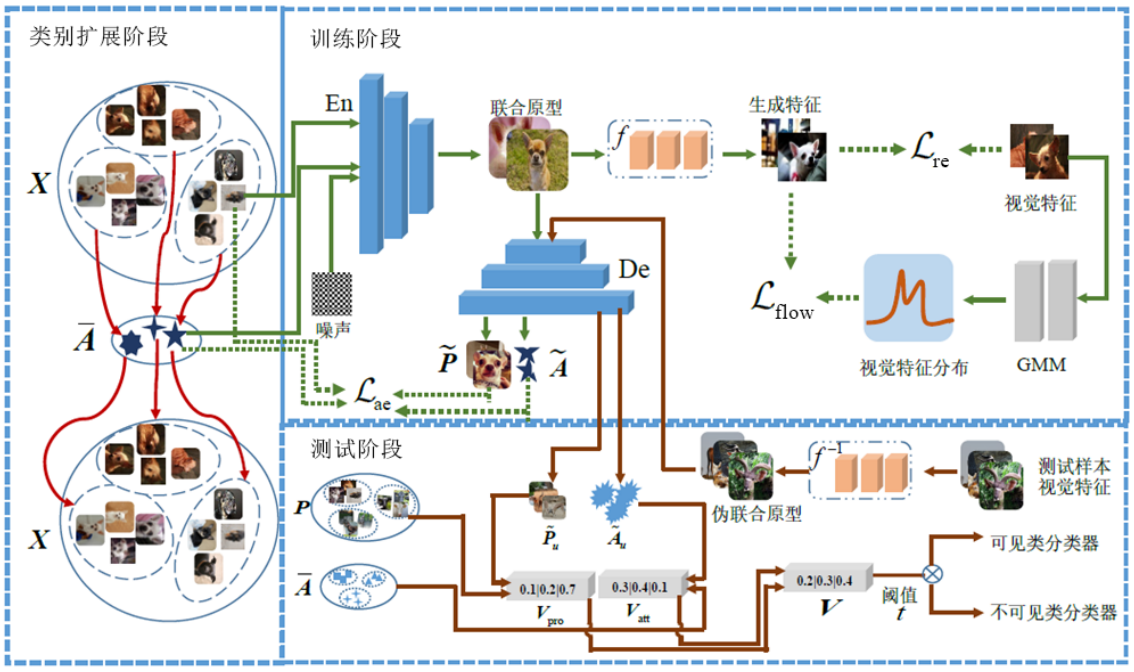


图 1　基于类别扩展的泛化零样本学习方法流程图

的类别标签，则对于每一个样本所属类别来说还存在着一个描述其视觉特征的 k 维语义属性向量 $\boldsymbol{a}_{y_i}$. 同样，对于包含 U 个类别共 n_u 个样本的不可见类集合，有 $\mathcal{T}_u=\left\{\boldsymbol{x}_i^u,y_i^u,\boldsymbol{a}_{y_i^u}^u\right\}$，其中，$\boldsymbol{x}_i^u\in\boldsymbol{X}_u^{n_u\times d}$，$y_i^u\in\{C+1,C+2,\cdots,C+U\}$，$\boldsymbol{a}_i^u\in\boldsymbol{A}_u^{U\times k}$，且 $\boldsymbol{x}_i^u$ 表示该集合中第 i 个样本的 d 维视觉特征向量，y_i^u 表示第 i 个样本的类别标签，$\boldsymbol{a}_{y_i^u}^u$ 表示第 i 个样本所属类别的 k 维语义属性信息.

对于广义零样本学习分类任务而言，本文在训练阶段仅能够获得可见类类别集合 $\mathcal{T}_s$ 以及不可见类集合中的类别语义属性信息 $\boldsymbol{A}_u$，而本文的目标则是希望在训练阶段中所得到的模型能够尽可能准确地将 $\boldsymbol{X}\cup\boldsymbol{X}_u$ 分类到 $C+U$ 维的类别空间中.

3.2　类扩展阶段

从 3.1 节的变量定义中可以看出，对于每一个类别来说，仅有一个 k 维的语义属性向量用来描述该类所有样本的视觉特征向量. 但是从现实场景出发，单一的语义属性可能无法完整地描述这个类别的视觉特征，例如对于狗这个类别来说，巴哥和藏獒有着巨大的视觉差异. 这种不充分的描述也将影响模型对各类别的描述能力，因此本文希望在每个类别上获得更多的语义属性向量来多样性地表达各类别在视觉空间上的差异.

为了达到上述目的，本文首先从视觉特征空间出发，通过 K-means[29] 聚类的方法将原始的单一类别划分为 N 个新类别，对于 N 数值的选择，将在 4.2 节中给出. 同时为了有足够的语义属性来描述这 N 个新类别，本文为原始的语义属性向量添加了 N 个可学习的随机高斯噪声来构建这 N 个新类别的语义属性向量. 设 $\boldsymbol{P}$ 为 $C\times N$ 个 d 维的视觉原型向量，$\boldsymbol{S}$ 表示 $C\times N$ 个 k 维的可学习随机高斯噪声，$\overline{\boldsymbol{A}}$ 表示对 $\boldsymbol{A}$ 的每一行复制 N 次后得到的 $(C\times N,k)$ 维语义属性. 此外，为了行文的简洁性，在不引起混淆的情况下不区分 $\overline{\boldsymbol{A}}$ 与 $\boldsymbol{A}$. 最后采用一个属性自编码器来学习这个 $\boldsymbol{S}$，具体的损失函数如式(1)所示：

$$\mathcal{L}_{\text{att}}=\left\|\boldsymbol{P}-\tilde{\boldsymbol{A}}\boldsymbol{W}\right\|^2+\lambda\left\|\boldsymbol{P}\boldsymbol{W}^{\mathrm{T}}-\tilde{\boldsymbol{A}}\right\|^2 \tag{1}$$

其中，$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{S}$，$\boldsymbol{W}\in\mathbb{R}^{k\times d}$ 表示一个从语义属性空间到视觉特征空间的映射函数，$\mathbb{R}^{k\times d}$ 表示一个 $k\times d$ 维的实数空间，λ 为一个超参数，$\|\cdot\|$ 则表示为一个 F 范数.

接下来对式(1)进行优化以求解语义噪声信息 $\boldsymbol{S}$，并以此为基础对不可见类进行类别的扩展. 容易知道式(1)同时关于变量 $\boldsymbol{S}$ 以及 $\boldsymbol{W}$ 是不可导的，但是对它们单独来说是一个可导函数. 因此本文将采用固定变量的方式对这两个参数进行迭代优化求解.

首先固定变量 $\boldsymbol{S}$ 为常量，则有式(1)关于变量 $\boldsymbol{W}$ 的偏导：

$$\frac{\partial\mathcal{L}_{\text{att}}}{\partial\boldsymbol{W}}=-2\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{P}+2\tilde{\boldsymbol{A}}^{\mathrm{T}}\tilde{\boldsymbol{A}}\boldsymbol{W}+2\lambda\boldsymbol{W}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}-2\lambda\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{P} \tag{2}$$

令式(2)为零，可以得到：

$$\tilde{\boldsymbol{A}}^{\mathrm{T}}\tilde{\boldsymbol{A}}\boldsymbol{W}+\lambda\boldsymbol{W}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}=(1+\lambda)\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{P} \tag{3}$$

式(3)是一个经典的 Sylvester 方程[30]，它有着如下所示的解析解：

$$\mathrm{vec}[\boldsymbol{W}]=\left(\boldsymbol{E}_d\otimes\tilde{\boldsymbol{A}}^{\mathrm{T}}\tilde{\boldsymbol{A}}+\lambda\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\otimes\boldsymbol{E}_k\right)^{-1}\mathrm{vec}\left[(1+\lambda)\tilde{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{P}\right] \tag{4}$$

其中，vec[•]表示对矩阵的列向量的堆叠，$\boldsymbol{E}_d$与$\boldsymbol{E}_k$分别表示$d\times d$维和$k\times k$维的单位矩阵，⊗表示一个Kronecker内积.

接下来，固定变量$\boldsymbol{W}$为常量，则可以得到式(1)关于$\boldsymbol{S}$的偏导如下所示：

$$\frac{\partial\mathcal{L}_{\mathrm{att}}}{\partial\boldsymbol{S}}=\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\boldsymbol{S}\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}-\boldsymbol{P}\boldsymbol{W}^{\mathrm{T}}+\lambda\boldsymbol{A}+\lambda\boldsymbol{S}-\lambda\boldsymbol{P}\boldsymbol{W}^{\mathrm{T}} \tag{5}$$

令式(5)为零并进行简单的化简可以得到：

$$\begin{aligned}&\boldsymbol{S}\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\lambda\boldsymbol{E}_s\right)=(1+\lambda)\boldsymbol{P}\boldsymbol{W}^{\mathrm{T}}-\mathbf{A}\mathbf{W}\boldsymbol{W}^{\mathrm{T}}-\lambda\boldsymbol{A}\\&\Rightarrow\boldsymbol{S}=(1+\lambda)\boldsymbol{P}\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}+\lambda\boldsymbol{E}_s\right)^{-1}-\boldsymbol{A}\end{aligned} \tag{6}$$

其中，$\boldsymbol{E}_s$表示一个$k\times k$维的单位矩阵.

由于不可见类与可见类有着相同的语义属性空间，因此本文可以在语义空间中为每一个不可见类别找到一个与其最为相似的可见类别，并将在该可见类别上学到的语义属性噪声施加到这个不可见类别上来，从而完成对不可见类别在语义空间上的类别扩展. 为此本文将在原始的语义属性空间上计算第i个不可见类别与全体可见类别之间的语义相似度的最大值：

$$I_i=\arg\max_j\mathrm{Simi}\left(\boldsymbol{a}_j,\boldsymbol{a}_i^u\right) \tag{7}$$

其中，$j=\{1,2,\cdots,C\}$，$\mathrm{Simi}(\cdot,\cdot)$表示一个相似度计算函数，本文使用余弦相似度函数来实现对语义相似度的计算.

通过式(7)将找到与不可见类最相似的可见类别. 另设$\tilde{\boldsymbol{a}}_i^u\in\mathcal{R}^{N\times k}$表示为对第$i$个不可见类进行扩展后的$N$个类别的语义属性向量，则有

$$\tilde{\boldsymbol{a}}_i^u=\overline{\boldsymbol{a}}_i^u+\overline{\boldsymbol{S}}_{I_i} \tag{8}$$

其中，$\overline{\boldsymbol{a}}_i^u$表示对$\boldsymbol{a}_i^u$进行$N$次堆叠，而$\overline{\boldsymbol{S}}_{I_i}=\left[\boldsymbol{S}_{(I_i-1)\times N+1},\boldsymbol{S}_{(I_i-1)\times N+2},\cdots,\boldsymbol{S}_{I_i\times N}\right]$. 最终将会得到扩展后的不可见类的语义属性向量矩阵$\tilde{\boldsymbol{A}}_u\in\mathcal{R}^{(U\times N,k)}$.

3.3 训练阶段

在这一节中，本文以生成流网络为基础来学习样本原型和视觉特征之间的映射关系. 由于生成流网络是具有可逆性的，这约束了它的输入及输出必须具有相同维度，因此本文还利用了编码器网络将样本的视觉特征原型以及语义属性原型映射到一个与样本视觉特征相同维度的原型空间中，之后再去寻找从该原型空间到视觉特征空间的变换关系. 设En是一个编码器网络，De是一个解码器网络，则有

$$\mathcal{L}_{\mathrm{ae}}=\left\|\left[\boldsymbol{P},\tilde{\boldsymbol{A}}\right]^{\mathrm{T}}-\mathrm{De}\left(\mathrm{En}\left(\boldsymbol{P},\tilde{\boldsymbol{A}},\boldsymbol{Z}\right)\right)\right\| \tag{9}$$

其中，$\boldsymbol{Z}$表示一个高斯噪声. 在通过En获得样本的原型变量后，本文将进一步学习一个模型ρ来将该原型变量映射到样本的真实视觉空间中去，然后通过最小化如下负对数似然函数对该模型进行优化：

$$\mathcal{L}_{\mathrm{flow}}=-\sum_{i=1}^{ns}\rho\left(\mathrm{En}\left(\boldsymbol{P}_{y_i},\tilde{\boldsymbol{a}}_{y_i},\boldsymbol{z}\right)\middle|\boldsymbol{x}_i\right) \tag{10}$$

其中，$\boldsymbol{P}_{y_i}$表示第i个样本所对应的视觉原型特征，$\boldsymbol{z}\in\mathcal{N}(\boldsymbol{0},\boldsymbol{1})$表示为一个多维高斯噪声. 设$f$表示一个可逆变换并利用该函数将En所输出的原型变量映射到视觉特征空间上，有$\boldsymbol{q}_i=f\left(\mathrm{En}\left(\boldsymbol{P}_{y_i},\tilde{\boldsymbol{a}}_{y_i},\boldsymbol{z}\right)\right)$，其中，变量$\boldsymbol{q}_i$的密度函数$\rho\left(\boldsymbol{q}_i|\boldsymbol{x}\right)$约束于样本的视觉特征，则根据变量替换理论可以得到：

$$\begin{aligned}&\log\rho\left(\mathrm{En}\left(\boldsymbol{P}_{y_i},\tilde{\boldsymbol{a}}_{y_i},\boldsymbol{z}\right)|\boldsymbol{x}_i\right)\\&=\log\rho\left(\boldsymbol{q}_i|\boldsymbol{x}_i\right)+\log\left(\left|\det\left(\frac{d_{\boldsymbol{q}_i}}{d_{\mathrm{En}\left(\boldsymbol{P}_{y_i},\tilde{\boldsymbol{a}}_{y_i},\boldsymbol{z}\right)}}\right)\right|\right)\end{aligned} \tag{11}$$

根据式(11)可以将式(10)改写为

$$\mathcal{L}_{\mathrm{flow}}=\log\left(\left|\det\left(\frac{d_{\boldsymbol{q}_i}}{d_{\mathrm{En}\left(\boldsymbol{P}_{y_i},\tilde{\boldsymbol{a}}_{y_i},\boldsymbol{z}\right)}}\right)\right|\right)-\sum_{i=1}^{n_s}\log\rho\left(f\left(\boldsymbol{q}_i\right)|\boldsymbol{x}_i\right) \tag{12}$$

式(12)的优化关键在于对变换f的选择上，通常来说我们需要保证对式(12)中第二项的计算足够简便并同时确保计算其逆变换的便捷性. 特别的是，如果变换f的雅可比行列式是一个上三角行列式，式(12)中的第二项将会简化为对行列式中对角线上的元素进行连乘. 由于在本文所提方法中不涉及对生成流网络的结构进行创新性的设计，因此本文不对生成流模块的网络结构以及变换函数的选择做过多的赘述，该网络模块的具体实现细节可以参考文献[12,31].

而为了保证由f所生成的伪样本能够更好地表示真实样本以帮助逆生成过程产生更真实的原型信息，进一步对$\boldsymbol{p}_i$做了如下所示约束：

$$\mathcal{L}_{\mathrm{re}}=\left\|\boldsymbol{p}_i-\boldsymbol{x}_i\right\| \tag{13}$$

可以得到最终的损失函数：

$$\mathcal{L}=\mathcal{L}_{\mathrm{ae}}+\alpha_1\mathcal{L}_{\mathrm{flow}}+\alpha_2\mathcal{L}_{\mathrm{re}} \tag{14}$$

其中，参数α_1与α_2用于平衡式(14)的各个组成部分.

3.4 分类阶段

3.4.1 预分类阶段

出于对可见类与不可见类别样本之间存在固有差异的考量，同时也为了能够缓解模型的偏置问题，本文利用预分类模块将测试集样本首先划分到可见类集合或不可见类集合中去. 当完成对3.3节中所述的生成流模型的训练后，本文将利用其逆变换为测试样本生成

伪原型变量,之后再利用解码器De将该伪原型变量解构为测试样本的伪语义原型与伪视觉原型. 当模型接收到测试样本$\boldsymbol{X}_u$时,有

$$\hat{\boldsymbol{A}}_u, \hat{\boldsymbol{P}}_u = \mathrm{De}\left(f^{-1}\left(\boldsymbol{X}_u\right)\right) \tag{15}$$

其中,$\hat{\boldsymbol{A}}_u$与$\hat{\boldsymbol{P}}_u$分别表示为测试样本生成的语义属性原型与视觉特征原型.

接下来,本文将分别计算测试样本的生成语义及生成视觉原型与可见类集中各类别的语义属性以及视觉原型之间的最大相似度得分,并将它们分别作为该测试样本在语义空间和视觉空间上的得分:

$$\begin{aligned} \boldsymbol{V}_{\mathrm{att}} &= \max\left(\mathrm{Simi}\left(\hat{\boldsymbol{A}}_u, \boldsymbol{S}\right)\right) \\ \boldsymbol{V}_{\mathrm{pro}} &= \max\left(\mathrm{Simi}\left(\hat{\boldsymbol{P}}_u, \boldsymbol{P}\right)\right) \end{aligned} \tag{16}$$

其中,$\boldsymbol{V}_{\mathrm{att}}$表示测试集样本的生成语义原型与可见类的语义属性原型之间的最大相似度得分,$\boldsymbol{V}_{\mathrm{pro}}$则表示测试样本的生成视觉原型与可见类的视觉原型之间的最大相似度得分,$\max(\cdot)$函数返回了测试样本的得分最大值.

由于视觉特征和语义属性对样本有着不同的辨识度,因此为了能够更好地区分可见类与不可见类样本,本文采用$\boldsymbol{V}_{\mathrm{att}}$与$\boldsymbol{V}_{\mathrm{pro}}$的加权得分作为最终的样本相似度得分. 对于测试集样本,其加权得分为

$$\boldsymbol{V} = a\boldsymbol{V}_{\mathrm{att}} + (1-a)\boldsymbol{V}_{\mathrm{pro}} \tag{17}$$

其中,加权系数$a \in [0,1]$,而$\boldsymbol{V}_{\mathrm{att}}$与$\boldsymbol{V}_{\mathrm{pro}}$均是通过余弦相似度函数计算得到的,因此其取值范围也均属于区间$[0,1]$,由此可知$\boldsymbol{V} \in [0,1]$.

出于同属于可见类集的样本将会有着更高的相似度得分的直观认识,在通过式(17)得到加权相似度得分$\boldsymbol{V}$后,本文将选定一个阈值$t \in (0,1)$来判断测试样本的归属问题. 对于那些加权得分大于阈值t的测试样本,将会把它划分到可见类集中去,否则将其分类到不可见类集合中去:

$$\hat{y}_i = \mathcal{I}\left(V_i > t\right) \tag{18}$$

其中,$\mathcal{I}(\cdot)$代表着一个指示函数,当$(\cdot)$为真时,其值为1,反之为0,即当$\hat{y}_i == 1$时,表示第i个测试样本被预分类到可见类集合中,否则将会被分类到不可见类集合中去.

3.4.2　正式分类阶段

经过式(18)后,完成了对测试样本的预分类处理. 之后根据可见类集与不可见类集的特点采用不同的分类器进行分类. 需要注意的是,本文将在经过类扩展的数据集上来为这个两个集合分别预训练不同的分类器. 但是由于类别扩展不会改变分类器模型的损失函数或者优化过程,同样也不会改变模型的网络结构,因此在本文中不再赘述如何训练将使用的Softmax分类器与CLSWGAN分类器.

由于在训练阶段我们能够关注到可见类样本,因此对于属于可见类集合中的测试样本,可以为它们在训练集上预训练一个监督分类模型来进行分类任务. 本文采用了经典的Softmax分类器:

$$\mathrm{lab}_u^j = \mathrm{Softmax}\left(\boldsymbol{X}_u^j\right), \quad \text{if} \quad \hat{y}_j == 1 \tag{19}$$

而对于那些预分类到不可见类集中的测试样本而言,初始的广义零样本问题退化为了传统零样本学习问题. 本文采用CLSWGAN模型来为不可见类集合中的样本进行分类:

$$\mathrm{lab}_u^j = \mathrm{CLSWGAN}\left(\boldsymbol{X}_u^j\right), \quad \text{if} \quad \hat{y}_j == 0 \tag{20}$$

结合式(19)与式(20),可以将得到各测试样本最终的预测标签$\tilde{y}_u^i$为

$$\tilde{y}_u^i = \begin{cases} \mathrm{lab}_s^i, & \hat{y}_i == 1 \\ \mathrm{lab}_u^i, & \hat{y}_i == 0 \end{cases} \tag{21}$$

4　实验

在这一节中,首先对广泛应用在泛化零样本分类学习任务中的四个数据集进行介绍,接着以对比表的形式报告本文所提方法与近些年来的一些经典方法在泛化零样本分类任务中的性能表现,此外还将以大量的实验来验证CEBGZSL方法的有效性.

4.1　数据集

在本文的实验中将会使用到两种粗粒度数据集以及三种细粒度数据集,对于这些数据集的相关信息的具体描述如下,同时表1对这些数据集的关键信息也进行了总结.

Animal With Attribute(AWA)[3]　作为一个关于动物的粗粒度数据集,AWA1数据集中共包含了50种类

表1　各数据集在PS划分方式下的详细信息

数据集	语义/视觉维度	训练样例数量	可见/不可见测试样例	可见/不可见类别	属性标注方式
AWA1	85/2048	19 832	5 685/4 958	40/10	专家标注
AWA2	85/2048	23 527	5 882/7 913	40/10	专家标注
CUB	312/2048	7 057	1 440/2 580	150/50	专家标注
SUN	102/2048	10 320	7 924/1 483	645/102	专家标注
FLO	1 024/2048	5 631	1 403/1 155	82/20	Word2Vec

注:表中数据均来自文献[33].

别共计30 475个样本. 其中有40个类别的样本将作为可见类参与模型的训练. 对于每一个类别来说,都有着一个85维的语义属性向量用来描述该类别的视觉特征.

Animal with Attribute2(AWA2)[32] 该数据集与AWA1数据集在类别分布上一致,它剔除了AWA1数据集中的一些无版权的图像样本,同时为每一个类别又添加了一些新的样本图像.

Caltech-UCSD Bird(CUB)[33] CUB数据集是一个有着11 788张鸟类图像的细粒度数据集. 它由200种不同类别但在视觉上十分接近的类别组成. 对于每一个类别而言有着一个312维的语义属性向量描述其视觉特征.

SUN Attribute(SUN)[34] SUN数据集同样是一个细粒度数据集. 它包含了来自717种不同类别的共计14 340个关于风景的图像样本,对于每一个类别,有102维的语义属性向量来描述它的视觉特征. 对于SUN数据集而言,每一个类别平均仅有约20个样本,这无疑会使得该数据集在包括图像分类任务在内的各种深度学习任务中成为一大挑战.

Oxford Flowers(FLO)[35] FLO数据集是一个由82种可见类与20种不可见类所组成的一个花卉集合细粒度数据集. 对于每一种类别而言都有着一个1 024维的语义属性用来描述它的视觉特征.

上述各个数据集中的语义属性向量通常可以通过两种方式获得,分别是基于专家标注的方法[36]和基于Word2Vec[37]的方法. 具体而言,基于专家标注的语义属性获取方法将首先由相关领域专家预定义若干种可能在所有类别中出现的视觉特征的描述,之后再对各个类别的描述进行0-1标注. 举例来说,对于动物数据集AWA1来说,首先由专家定义了85种可能出现的视觉特征描述,然后具体到狗这个类别上,在例如“有皮毛”“有尾巴”等维度上标注为“1”,而在例如“像马的”“有鳞片的”等维度上标注为“0”. 但是当待标注数据集是超大型数据集或细粒度数据集时,采用专家标注的方式将会是一件极其困难的事情. 此时第二种基于神经网络的Word2Vec方法将是一个性价比极高的语义属性获取方式. 对于数据集中的每一个类别,可以使用它的文本资料(例如维基百科中的相关词条)作为自然语言处理模型的语料库来自动学习该类别对应的语义属性向量.

4.2 实验参数设置

本节将对文中所涉及的相关变量的取值或选择方法做进一步的说明. 对于第3.2节中所提到的类别扩展的倍数N,我们在所有数据集上均设置为3. 对于式(14)中的参数α_1与α_2的选择,在粗粒度数据集上均分别根据经验设置为0.1与0.001,而在细粒度数据集上则均设置为0.001与0.01. 对解码器与编码器来说,都表示一个隐藏层节点数为4 096,采用LeakyRelu激活函数的感知机网络. 对于本文的方法模型中所存在的其他超参数来说,往往不同的取值会给模型带来不同的性能. 此外对于不同的数据集而言,由于数据本身存在的差异以及数据在分布上存在的不同而存在着不同的最优参数,因此本文将采用交叉验证的方式来寻找最优的参数. 由于在零样本学习的设置中我们无法得到不可见类的相关信息,因此依次将部分的可见类视为不可见类用作验证集. 具体而言,随机选择20%的可见类别视作不可见类用于验证,然后选择那些能够使模型取得最佳平均实验结果的参数作为最终的模型参数,同时这些参数也将应用到对不可见类进行分类预测中去.

4.3 广义零样本分类实验

本节在表2中报告并分析CEBGZSL在广义零样本分类任务中的性能表现. 但需注意的是,由于在GZSL的设置中模型无法在训练阶段接触到任何不可见类的视觉特征信息,因此模型会更倾向将属于不可见类的测试样本分类到可见类中. 反映到分类准确率上则会出现模型会更容易得到一个较高的可见类平均分类准确率Se与一个较低的不可见类平均分类准确率Un. 因此为了能够更好地反应模型的性能,同时也为了能够更好地与其他方法进行对比,本文同样遵循了Xian等[32]的建议,即使用Se与Un之间的调和平均准确率$H=2\times \mathrm{Se}\times \mathrm{Un}/(\mathrm{Se}+\mathrm{Un})$作为最终衡量模型性能的指标. 表2对比报告了CEBGZSL与一些优秀零样本学习方法在图像分类任务中的表现.

从表2中可以看出本文所提方法在全部数据集上都能取得更好的分类表现. 在AWA1数据集的调和平均准确率上,有着66.5%分类准确率的次优方法RFF-GZSL仍比本文方法低1.7%. 同样,在AWA2数据集上,CEBGZSL的性能相对于次优方法IZF-NBC提升了2.8个百分点,达到了68.7%.

在细粒度数据集CUB,SUN以及FLO上,CEBGZSL同样都有着很好的分类效果. 在CUB数据集上,次优方法RFF-GZSL有着54.6%的分类表现,低于本文方法2.6个百分点. 在SUN数据集上CEBGZSL比次优方法IEF-NBC高出了2.3个百分点,相比较于表中所列的分类效果最差的SZSL方法提升了10个百分点. 在FLO数据集上,本文方法相对于次优方法RFF-GZSL则提高了2.1个百分点.

此外,从表2中还能观察到偏置问题对模型在广义零样本分类任务中的严重影响. DUET方法在AWA1以及AWA2数据集上都有着高达90%的可见类平均分类

表2　在AWA1,AWA2,CUB,SUN与FLO数据集上的广义零样本分类表现

方法	数据集														
	AWA1			AWA2			CUB			SUN			FLO		
	Se	Un	*H*	Se	Un	*H*	Se	Un	*H*	Se	Un	*H*	Se	Un	*H*
CLSWGAN[18]	57.9	61.4	59.6	—	—	—	43.7	57.7	49.7	42.6	36.6	39.4	59.0	**73.8**	65.6
DUET[38]	**90.1**	47.5	62.2	**90.2**	48.2	63.4	80.1	39.7	53.1	—	—	—	—	—	—
COSMO[27]	80.0	52.8	63.6	—	—	—	**87.8**	44.4	50.2	37.7	44.9	41.0	81.4	59.6	68.8
RFF-GZSL[39]	75.1	59.8	66.5	—	—	—	56.6	52.6	54.6	38.6	45.7	41.9	78.2	65.2	71.1
LsrGAN[40]	74.6	54.6	63.0	—	—	—	59.1	48.1	53.0	37.7	44.8	40.9	—	—	—
DRN[41]	81.4	50.1	62.1	85.3	44.9	58.8	58.8	46.9	52.2	—	—	—	—	—	—
OBTL[42]	—	—	—	73.4	59.5	65.7	59.9	44.8	51.3	42.9	44.8	43.8	—	—	—
IZF-NBC[14]	75.2	57.8	65.4	76.0	58.1	65.9	56.3	44.2	49.5	**50.6**	44.5	47.4	—	—	—
GZSL-DR[43]	72.9	60.7	66.2	80.2	56.9	66.6	58.2	51.1	54.4	47.6	36.6	41.4	—	—	—
CEBGZSL	75.0	**62.6**	**68.2**	78.3	**61.2**	**68.7**	60.9	**53.3**	**56.9**	44.6	**56.3**	**49.7**	**90.6**	61.3	**73.2**

注:表中其他方法的实验数据均来自对应文献中的实验结果,对于原文中没有的实验数据,使用'—'进行标注,对于每一列的最大值,使用加粗标记.

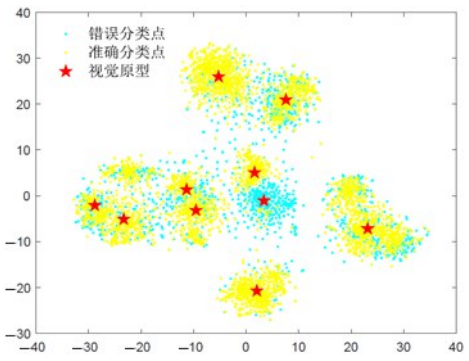
准确率,在CUB数据集上也有着高达80%的可见类平均分类准确率.但是因为该方法在不可见类上的表现较差使得其调和平均分类准确率较低,在AWA1数据集上仅比最差的方法高出了0.1%.而造成这一现象的根本原因在于该方法没有对模型的偏置问题进行处理,从而造成测试集中的样本更容易被分类到可见类集中去.而CEBGZSL从样本出发,使用预分类的方式缩小待分类样本的分类空间,从而缓解了模型容易误分不可见类样本到可见类集上的问题,最终使得模型在可见类与不可见类集上有着更加均衡的性能表现.

同时,为了更加直观地展示类别扩展方法对本文模型有着怎么样的影响,进一步通过图2(*a*)与图2(*b*)展示了模型在有无类别扩展帮助下对AWA1数据集中的不可见类样本的分类能力.从图2(*a*)与图2(*b*)的对比中可以观察到,在两图的正上方、正下方及右下方区域被错误分类的样本的数量有着明显的减少.造成这一现象的原因在于,当通过类别扩展的方法为每个不可见类别得到更多的原型样本后,这些原型样本可以有效地帮助模型识别测试样本的所属类别.

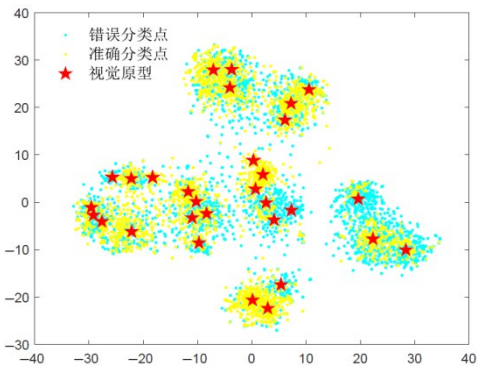
最后,本文还将利用图3展示在CUB数据集上不同的属性标注方式对模型分类性能的影响.图3中三组柱状图的左侧反斜线柱体表示基于Word2vec提取的语义属性下的模型性能表现,右侧砖状柱体则表示模型基于专家标注下的性能表现.正如图3所示,由Word2vec方式提取的属性帮助模型在三组下都取得了更好地表现.造成这一现象的根本原因是由网络提取到的复杂的语义属性相对于专家标注的语义属性更加精细,更能有效地区分不同类别之间的细微差别.

4.4　消融实验

从第3节中的方法论可以看出,在原始数据集经过



(*a*) CLSWGAN[18]



(*b*) CEBGZSL

图2　基于t-SNE所展示的模型在AWA1数据集上的分类效果

类别扩展处理之后,本文方法将经过三个主要模块来完成对模型的训练.本节将讨论自编码器模块,流生成网络模块以及重建模块这三个主要模块对整个模型的性能影响.表3展示了各模块在AWA2数据集以及CUB数据集上对模型分类性能的影响.

从表3可以看出,$\mathcal{L}_{\text{flow}}$与$\mathcal{L}_{\text{re}}$对模型的性能在两个数据集上均有很大影响.具体来看,当不使用生成流模

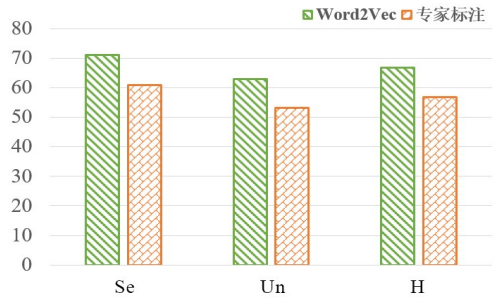


图3 不同语义属性对模型在CUB数据集上的分类表现

表3 各模块对模型分类性能的影响

消融模块	数据集					
	AWA2			CUB		
	Se	Un	H	Se	Un	H
$\mathcal{L}$	78.29	61.21	68.71	60.89	53.32	56.85
w\o $\mathcal{L}_{\text{flow}}$	66.66	50.37	57.38	54.07	47.49	50.56
w\o $\mathcal{L}_{\text{re}}$	65.99	56.05	60.61	54.53	49.65	51.98
w\o$\mathcal{L}_{\text{flow}}+\mathcal{L}_{\text{re}}$	58.06	42.04	48.77	47.20	44.43	45.78

注:w\o表示不使用该损失.

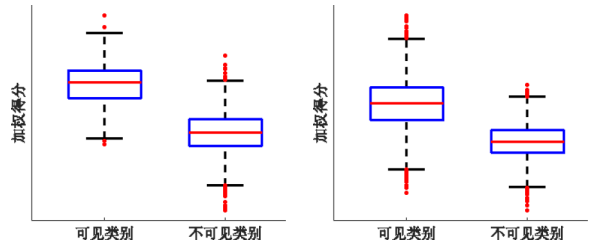
块,直接利用自编码器模块完成从样本原型到样本视觉特征之间的映射,同时约束编码器输出的原型特征与真实视觉特征是十分接近的时候,模型的分类性能表现在AWA2数据集上下降了约10%,在CUB数据集上下降了约6%. 并且当进一步放开对原型特征的约束即$\mathcal{L}_{\text{re}}$损失时,模型的分类性能将会进一步大幅下降,在两个数据集上分别下降了约9%和5%.

而当仅放开该约束时,模型的性能在AWA2数据集上将会下降8.1%,在CUB数据集模型下降约5%. 由以上分析看出,生成流模块能够帮助我们更好地寻找两个分布之间的映射关系,从而帮助测试集样本生成更具有辨别力的语义原型及视觉原型. 而$\mathcal{L}_{\text{re}}$损失则能够使得模型生成更加真实的视觉特征并进一步提高解码器网络的解耦能力.

4.5 预分类结果分析

为能够在一定程度上缓解模型的偏置问题,同时也为能够更好地对数据本身的固有差异加以利用,本文所提方法中采用一个预分类模块对测试集样本进行可见类-不可见类的二分类处理. 本节将对由式(17)计算到的样本加权得分分别从分布以及数值上进行直观的展示.

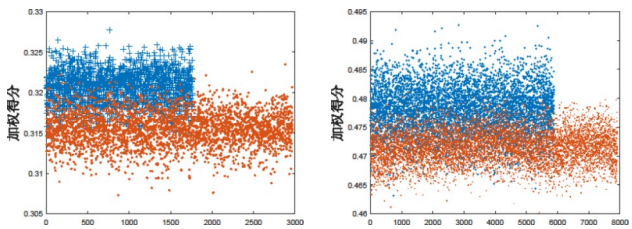
对于图4中所示的箱体图,箱体中间的红色横线表示数据的中位数,上下边界则分别表示着数据分布的上下四分位数,此外箱体上下的两条横线则分别代表着数据分布的两个极值,红色标记则表示处于分布外的离散值. 容易观察到,图4(a)和图4(b)分别所示的AWA2与CUB数据集上的测试集得分有着相似的分布,即在可见类与不可见类的得分分布上均有着较大的差异. 在两图中均有着不可见类别的极大值与可见类的上四分位数较为接近,由此可以得出这样的结论:至少能够保证75%的样本准确地分类到它所属的集合中去.



(a) AWA2数据集 (b) CUB数据集

图4 预分类得分分布可视化

而对于图5中所示的散点图,图像的纵轴表示加权得分,横轴表示全部的测试样本,并且蓝色"+"符号表示可见类样本得分,而红色的"·"符号则表示不可见类的样本得分. 从图5(a)与图5(b)的对比中可以更加直观地看出,在可见类与不可见类样本的加权得分上存在着明显的分界线. 结合图4可以看出本文中预分类模块的有效性.



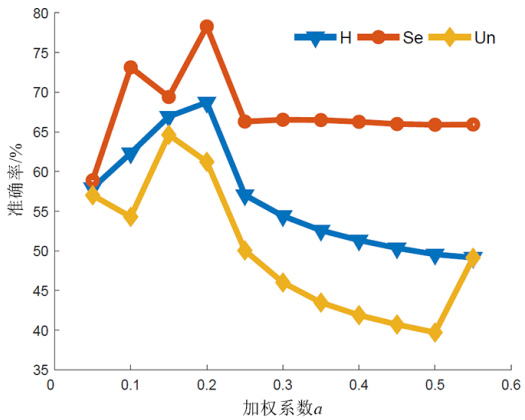
(a) AWA2数据集 (b) CUB数据集

图5 预分类得分数值可视化

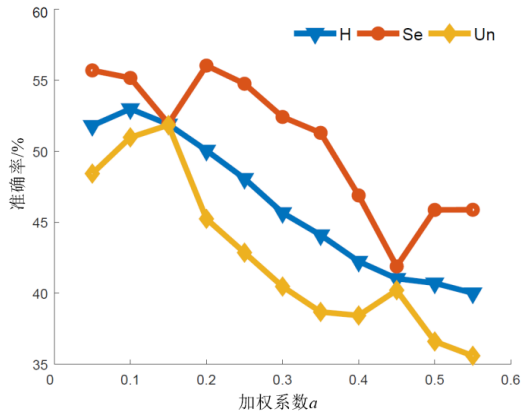
4.6 加权系数对模型的影响

在接收到测试集样本后,模型最终将同时生成视觉原型与语义原型并在预分类模块中以这两个原型为依据并根据式(17)计算该测试样本的加权得分来进行二分类. 本节将对式(17)中的加权系数a的取值,即语义属性和视觉特征对分类结果的影响程度进行讨论.

图6(a)与图6(b)分别展示了加权系数在AWA2数据集和CUB数据集上影响. 由式(17)可知,a控制着语义属性对加权得分$\boldsymbol{V}$的影响程度. 当$a=0$时,$\boldsymbol{V}$完全取决于视觉特征的影响;当$a=1$时,$\boldsymbol{V}$完全取决于语义属性的影响. 从图6(a)及图6(b)可以看出,无论当$\boldsymbol{V}$完全依赖语义属性还是视觉特征,均不能保证模型能够取得最好的分类表现. 在AWA2和CUB数据集上,模型的性能均更多地受到视觉特征的影响,并且当加权系数在[0.1,0.3]之间时,模型的性能有着更好的表现.



(a) AWA2数据集



(b) CUB数据集

图6　加权系数a对准确率影响的分析

4.7　类扩展分析

原始的数据集上存在着一些类别在视觉空间上有着较大的类内差异但在语义空间上却共享唯一的语义属性的不合理性. 为此,本文对各类别进行了扩展,同时为扩展后的各类别学习对应的语义属性向量. 本节将针对类扩展部分进行讨论与分析.

图7展示了四种数据集的可见类别在经过类别扩展前后的平均类内差异,图中最左侧蓝色柱体与中间的橙色柱体分别表示扩展前和扩展后的平均类内距离,最右侧的针状图则展示了扩展前后平均类内差异的比例. 从图中可以看出,除在CUB数据集外,其余数据集的平均类内距离均减小了90%以上,在SUN数据集上甚至达到了95%. 对于CUB数据集,在经过类别扩展后,数据集的平均类内差异也减小了约80%.

回到我们的出发点,我们期望对每一个类别而言都有足够的语义属性来进行描述. 因此当我们对各类别在视觉空间上完成扩展后,将继续通过式(1)实现在

图7　类扩展前后可见类别的平均类内差异及其变化率

语义空间上对类别的扩展,最后通过式(8)实现对不可见类进行扩展. 在上一部分我们验证了扩展后的各类别在视觉空间上有着更小的类内差异,在这一部分将进一步验证扩展后的语义属性的有效性. 设$\boldsymbol{M}$表示从语义空间到视觉空间的映射函数,则有$\boldsymbol{M}=\boldsymbol{XA}^{-1}$. 接着利用$\boldsymbol{M}$将不可见类的语义属性映射到视觉空间上以得到各类别的视觉原型并通过计算各类别的真实视觉特征到相应的视觉原型的平均距离作为衡量标准. 本文在粗粒度数据集和细粒度数据集中间分别选择了AWA2与CUB进行实验并在表4中展示了实验结果,表中的扩展后(V1)与扩展后(V2)分别表示使用余弦相似度和欧氏距离作为度量函数.

表4　类别扩展前在语义属性空间上的有效性分析

数据集	扩展前	扩展后(V1)	扩展后(V2)
AWA2	53.08	39.97	39.98
CUB	40.36	30.37	30.39

表4中的数据反映了数据集中各类别到伪视觉原型距离的平均值,其值越小表示着该伪视觉原型更加能够代表这个类别. 从表4中可以看出,无论是在粗粒度数据集AWA2上还是在细粒度数据集CUB上,在经过类别扩展后均有着更小的结果. 由此可知,相比经过类别扩展后的多语义属性信息而言,单一的语义属性确实不能很好地用来描述各个类别. 结合图7来看,对类别进行扩展无论是在语义空间还是在视觉空间,都能更好地反应出类别内部的差异性,采用更多的语义对这个差异进行描述也能够更好地帮助模型辨别各类别.

而在选择式(7)中的相似度函时,本文选择了余弦相似度函数来衡量不可见类与可见类的语义属性之间的相似性. 这是由于在衡量语义属性向量之间的相似性时,我们更希望在向量空间上保持相似,而非在数值上保持着大小的相似性. 从表4中可以看出,式(7)中的相似度函数的选择对我们的类别扩展的思想并没有太大的影响. 此外,还可以看出,第三列中的结果是略优于第四列的,这也在一定程度上反映了本文选择在向量方向上保持相似性的有效性.

5 结论

为缓解各类别中单一的语义属性无法很好描述类别内存在较大视觉差异的问题,本文对原始的类别通过聚类的方式进行了扩展并同时利用自编码器的方法为每一个新类别学习新的语义属性,最终在原始数据集中的视觉空间上对各类别内部进行了细化,同时也在语义空间上学习了更加丰富的属性变量来充分描述这些细化后的类内差异. 出于对数据集中可见类与不可见类在分布上具有天然差异的认知以及生成流模型能够较好拟合两个任意分布的特点,本文以生成流网络为基础构建了一种以预分类结果为基础的广义零样本分类模型. 最终通过大量的实验表明了本文所提方法的有效性和优越性.

参考文献

[1] LAROCHELLE H, ERHAN D, BENGIO Y. Zero-data learning of new tasks[C]//Proceedings of the 23rd national conference on Artificial intelligence-Volume 2. Chicago: AAAI Press, 2008: 646-651.

[2] 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[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 52-68.

[3] LAMPERT C H, NICKISCH H, HARMELING S. Attribute-based classification for zero-shot visual object categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 453-465.

[4] 程玉虎, 乔雪, 王雪松. 基于混合属性的零样本图像分类[J]. 电子学报, 2017, 45(6): 1462-1468.
CHENG Y H, QIAO X, WANG X S. Hybrid attribute-based zero-shot image classification[J]. Acta Electronica Sinica, 2017, 45(6): 1462-1468. (in Chinese)

[5] 赵鹏, 汪纯燕, 张思颖, 等. 一种基于融合重构的子空间学习的零样本图像分类方法[J]. 计算机学报, 2021, 44(2): 409-421.
ZHAO P, WANG C Y, ZHANG S Y, et al. A zero-shot image classification method based on subspace learning with the fusion of reconstruction[J]. Chinese Journal of Computers,2021, 44(2): 409-421. (in Chinese)

[6] BAI H Y, ZHANG H F, WANG Q. Dual discriminative auto-encoder network for zero shot image recognition[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(3): 5159-5170.

[7] ZHANG H F, LIU L, LONG Y, et al. Deep transductive network for generalized zero shot learning[J]. Pattern Recognition, 2020, 105: 107370.

[8] 冀中, 汪浩然, 于云龙, 等. 零样本图像分类综述: 十年进展[J]. 中国科学: 信息科学, 2019, 49(10): 1299-1320.
JI Z, WANG H R, YU Y L, et al. A decadal survey of zero-shot image classification[J]. Scientia Sinica Informationis, 2019, 49(10): 1299-1320. (in Chinese)

[9] KODIROV E, XIANG T, GONG S G. Semantic autoencoder for zero-shot learning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4447-4456.

[10] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[11] KINGMA D P, WELLIING M. Auto-encoding variational bayes[C]//In Proceedings of the International Conference on Learning Representations. Banff: ICLR, 2014: 1-14.

[12] DINH L, KRUEGER D, BENGIO Y. Nice: Non-linear independent components estimation[C]//Proceedings of the International Conference on Learning Representations. San Diego: ICLR, 2015: 1-14.

[13] CHEN Z, LUO Y D, WANG S, et al. Mitigating generation shifts for generalized zero-shot learning[C]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Conference: ACM, 2021: 844-852.

[14] SHEN Y M, QIN J, HUANG L, et al. Invertible zero-shot recognition flows[C]//European Conference on Computer Vision. Virtual Conference: Springer, 2020: 614-631.

[15] LIU J R, FU L Y, ZHANG H F, et al. Learning discriminative and representative feature with cascade GAN for generalized zero-shot learning[J]. Knowledge-Based Systems, 2022, 236: 107780.

[16] ZHANG H F, WANG Y D, LONG Y, et al. Modality independent adversarial network for generalized zero shot image classification[J]. Neural Networks: The Official Journal of the International Neural Network Society, 2021, 134: 11-22.

[17] ZHANG H F, BAI H Y, LONG Y, et al. A plug-in attribute correction module for generalized zero-shot learning [J]. Pattern Recognition, 2021, 112: 107767.

[18] XIAN Y Q, LORENZ T, SCHIELE B, et al. Feature generating networks for zero-shot learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5542-5551.

[19] SARIYILDIZ M B, CINBIS R G. Gradient matching generative networks for zero-shot learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recog-

nition. Long Beach: IEEE, 2019: 2163-2173.

[20] SCHÖNFELD E, EBRAHIMI S, SINHA S, et al. Generalized zero- and few-shot learning via aligned variational autoencoders[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8239-8247.

[21] WANG W L, PU Y C, VERMA V, et al. Zero-shot learning via class-conditioned deep generative models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018, 32(1): 4211-4218.

[22] ARDIZZONE L, KRUSE J, ROTHER C, et al. Analyzing inverse problems with invertible neural networks[C]//Proceedings of the International Conference on Learning Representations. Vancouver: ICLR, 2018: 1-20

[23] 王格格, 郭涛, 余游, 等. 基于生成对抗网络的无监督域适应分类模型[J]. 电子学报, 2020, 48(6): 1190-1197.
WANG G G, GUO T, YU Y, et al. Unsupervised domain adaptation classification model based on generative adversarial network[J]. Acta Electronica Sinica, 2020, 48(6): 1190-1197. (in Chinese)

[24] 席亮, 刘涵, 樊好义, 等. 基于深度对抗学习潜在表示分布的异常检测模型[J]. 电子学报, 2021, 49(7): 1257-1265.
XI L, LIU H, FAN H Y, et al. Deep adversarial learning latent representation distribution model for anomaly detection[J]. Acta Electronica Sinica, 2021, 49(7): 1257-1265. (in Chinese)

[25] CHEN X Y, LAN X G, SUN F C, et al. A boundary based out-of-distribution classifier for generalized zero-shot learning[C]//European Conference on Computer Vision. Virtual Conference: Springer, 2020: 572-588.

[26] DAVIDSON T R, FALORSI L, DE CAO N, et al. Hyperspherical variational auto-encoders[C]//Proceeding of the Conference on Uncertainty in Artificial Intelligence 2018. California: AUAI, 2018: 856-865.

[27] ATZMON Y, CHECHIK G. Adaptive confidence smoothing for generalized zero-shot learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 11663-11672.

[28] MIN S B, YAO H T, XIE H T, et al. Domain-aware visual bias eliminating for generalized zero-shot learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 12661-12670.

[29] QUEEN J MAC. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. California: The Regents of the University of California, 1967: 281-297.

[30] BARTELS R H, STEWART G W. Solution of the matrix equation AX + XB = C [F_4][J]. Communications of the ACM, 1972, 15(9): 820-826.

[31] KINGMA D P, DHARIWAL P. Glow: Generative flow with invertible 1×1 convolutions[C]//Proceedings of the Anuual Conference on Neural Information Processing Systems. Montreal: NIPS, 2018: 10236-10245.

[32] XIAN Y Q, LAMPERT C H, SCHIELE B, et al. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2251-2265.

[33] WELINDER P, BRANSON S, MITA T, et al. The caltech-ucsd birds-200-2011 dataset[R]. California: California Institute of Technology, Technical Report: CNS-TR-2010-001, 2010.

[34] PATTERSON G, XU C, SU H, et al. The SUN attribute database: Beyond categories for deeper scene understanding[J]. International Journal of Computer Vision, 2014, 108(1/2): 59-81.

[35] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes[C]//2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar: IEEE, 2008: 722-729.

[36] LAMPERT C H, NICKISCH H, HARMELING S. Learning to detect unseen object classes by between-class attribute transfer[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 951-958.

[37] CHURCH K W. WORD2VEC[J]. Natural Language Engineerring, 2017, 23(1): 155-162.

[38] JIA Z, ZHANG Z, WANG L, et al. Deep unbiased embedding transfer for zero-shot learning[J]. IEEE Transactions on Image Processing, 2019, 29: 1958-1971.

[39] HAN Z Y, FU Z Y, YANG J. Learning the redundancy-free features for generalized zero-shot object recognition [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 12862-12871.

[40] VYAS M R, VENKATESWARA H, PANCHANATHAN S. Leveraging seen and unseen semantic relationships for generative zero-shot learning[C]//European Conference on Computer Vision. Virtual Conference: Spring-

er, 2020: 70-86.
[41] HUANG S, LIN J K, HUANGFU L W. Class-prototype discriminative network for generalized zero-shot learning [J]. IEEE Signal Processing Letters, 2020, 27: 301-305.
[42] KESHARI R, SINGH R, VATSA M. Generalized zero-shot learning via over-complete distribution[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 13297-13305.
[43] LI X Y, XU Z, WEI K, et al. Generalized zero-shot learning via disentangled representation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Conference: AAAI Press, 2021, 35(3): 1966-1974.

作者简介

张 杰 男,1997年出生于安徽省滁州市.南京理工大学计算机科学与工程学院硕士研究生.主要研究方向为零样本学习.

廖盛斌 男,1969年出生于湖北省公安市.华中师范大学国家数字化学习工程技术研究中心副教授、博士生导师.主要研究方向为大数据与机器智能、智慧教育、最优化理论与算法等.

张浩峰(通讯作者) 男,1983年出生于江苏省淮安市.南京理工大学计算机科学与工程学院教授、博士生导师.主要研究方向为深度学习理论及应用、多媒体数据检索与分类等.
E-mail: zhanghf@njust.edu.cn

陈得宝 男,1975年出生于安徽省安庆市.淮北师范大学计算机科学与技术学院院长,教授、硕士生导师.主要研究方向为进化计算与智能优化.