一种基于融合重构的子空间学习的零样本图像分类方法

赵 鹏[1),2)] 汪纯燕[2)] 张思颖[2)] 刘政怡[1),2)]

[1)](安徽大学计算智能与信号处理教育部重点实验室 合肥 230601)

[2)](安徽大学计算机科学与技术学院 合肥 230601)

摘 要 图像分类是计算机视觉中一个重要的研究子领域.传统的图像分类只能对训练集中出现过的类别样本进行分类.然而现实应用中,新的类别不断涌现,因而需要收集大量新类别带标记的数据,并重新训练分类器.与传统的图像分类方法不同,零样本图像分类能够对训练过程中没有见过的类别的样本进行识别,近年来受到了广泛的关注.零样本图像分类通过语义空间建立起已见类别和未见类别之间的关系,实现知识的迁移,进而完成对训练过程中没有见过的类别样本进行分类.现有的零样本图像分类方法主要是根据已见类别的视觉特征和语义特征,学习从视觉空间到语义空间的映射函数,然后利用学习好的映射函数,将未见类别的视觉特征映射到语义空间,最后在语义空间中用最近邻的方法实现对未见类别的分类.但是由于已见类和未见类的类别差异,以及图像的分布不同,从而容易导致域偏移问题.同时直接学习图像视觉空间到语义空间的映射会导致信息损失问题.为解决零样本图像分类知识迁移过程中的信息损失以及域偏移的问题,本文提出了一种图像分类中基于子空间学习和重构的零样本分类方法.该方法在零样本训练学习阶段,充分利用未见类别已知的信息,来减少域偏移,首先将语义空间中的已见类别和未见类别之间的关系迁移到视觉空间中,学习获得未见类别视觉特征原型.然后根据包含已见类别和未见类别在内的所有类别的视觉特征原型所在的视觉空间和语义特征原型所在的语义空间,学习获得一个潜在类别原型特征空间,并在该潜在子空间中对齐视觉特征和语义特征,使得所有类别在潜在子空间中的表示既包含视觉空间下的可分辨性信息,又包含语义空间下的类别关系信息,同时在子空间的学习过程中利用重构约束,减少信息损失,同时也缓解了域偏移问题.最后零样本分类识别阶段,在不同的空间下根据最近邻算法对未见类别样本图像进行分类.本文的主要贡献在于:一是通过对语义空间中类别间关系的迁移,学习获得视觉空间中未见类别的类别原型,使得在训练过程中充分利用未见类别的信息,一定程度上缓解域偏移问题.二是通过学习一个共享的潜在子空间,该子空间既包含了图像视觉空间中丰富的判别性信息,也包含了语义空间中的类别间关系信息,同时在子空间学习过程中,通过重构,缓解知识迁移过程中信息损失的问题.本文在四个公开的零样本分类数据集上进行对比实验,实验结果表明本文提出的零样本分类方法取得了较高的分类平均准确率,证明了本文方法的有效性.

关键词 零样本图像分类;迁移学习;子空间学习;重构;特征原型

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2021.00409

A Zero-Shot Image Classification Method Based on Subspace Learning with the Fusion of Reconstruction

ZHAO Peng[1),2)] WANG Chun-Yan[2)] ZHANG Si-Ying[2)] LIU Zheng-Yi[1),2)]

[1)](*Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei* 230601)

[2)](*School of Computer Science and Technology, Anhui University, Hefei* 230601)

Abstract Image classification is an important research subfield in the computer vision. Traditional

收稿日期:2019-08-18;在线发布日期:2020-05-05.本课题得到国家自然科学基金(61602004)、安徽省高校自然科学研究重点项目(KJ2018A0013,KJ2017A011)、安徽省自然科学基金(1908085MF188,1908085MF182)、安徽省重点研究与开发计划项目(1804d08020309)资助.**赵 鹏**,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究方向为机器学习、图像理解.E-mail:zhaopeng_ad@163.com.**汪纯燕**,硕士研究生,主要研究方向为机器学习、图像分类.**张思颖**,硕士研究生,主要研究方向为机器学习、图像分类.**刘政怡**,博士,副教授,硕士生导师,中国计算机学会(CCF)会员,主要研究方向为机器学习、计算机视觉.

image classification can only classify the samples of the seen categories which have appeared in the training dataset. However, new categories continue to emerge in real-world applications. The samples of the new categories should be collected and the classifier should be retrained. Different from traditional classification methods, zero-shot image classification aims at classifying the samples of the unseen categories which have not appeared in the training dataset. Zero-shot classification is a very challenging task and has attracted much attention in recent years. Zero-shot image classification bridges the seen categories and the unseen categories through the semantic embedding space, which transfers knowledge from the seen categories to the unseen categories and classifies the samples from the unseen categories. Firstly, the existing zero-shot classification methods typically learn a mapping function from the visual space to the semantic embedding space only according to the information of the samples from the training seen categories. Then, the learned mapping function is utilized to map the visual feature of the test sample from the unseen categories to the semantic space. Finally, zero-shot recognition classify the test samples from the unseen categories by a simple nearest neighbor search in the semantic embedding space. But the seen categories and the unseen categories are different, which will lead to the domain shift. Moreover, directly learning the mapping function from visual space to semantic embedding space will lead to the information loss. In order to solve the problems of the information loss and the domain shift in the knowledge transfer of zero-shot image classification, we propose a zero-shot classification approach based on subspace learning and reconstruction for image classification (Zero-Shot Classification based on Subspace learning and Reconstruction, ZSCSR). Firstly, ZSCSR makes full use of the unseen category information to mitigate the domain shift problem. It transfers the relationship between the seen categories and the unseen categories from the semantic embedding space into the visual space, and obtains the visual prototypes of the unseen categories. Then, according to the visual prototypes and semantic prototypes of all categories including the seen and the unseen categories, ZSCSR learns a latent subspace, which aligns the visual and the semantic spaces. The latent subspace not only contains the discriminative information in the visual space, but also contains the information of the category relationships in the semantic embedding space. Meanwhile, the reconstruction constraint reduces the information loss in the subspace learning. Finally, in the zero-shot recognition, the test samples of unseen classes could be classified by the nearest neighbor search in different spaces. There are two main contributions in this paper as follows. (1) ZSCSR learns the visual prototype of the unseen categories through transferring the relationship between the seen categories and the unseen categories from the semantic embedding space to the visual space, which relieves the domain shift problem. (2) ZSCSR learns a latent space through the latent space learning and reconstruction, which reduces the information loss. The proposed method is evaluated for zero-shot recognition on four benchmark datasets. The experimental results show the proposed method achieves higher average accuracies, which prove the effectiveness of the proposed method.

Keywords zero-shot image classification; transfer learning; subspace learning; reconstruction; feature prototype

1 引 言

传统的分类方法只能对训练数据集中出现过的类别样本进行分类,却无法对训练数据集中未出现过的类别样本进行分类.然而现实场景中,新类别往往层出不穷.收集足够数量的新类别标注样本通常费时费力,在某些特定领域甚至无法获取足够数量的

新类别标注样本. 零样本学习(Zero-Shot Learning, ZSL)应运而生,受到越来越多的研究者的关注. 人类具备识别未见类别样本的能力[1],例如一个只见过马而未见过斑马的孩子,如果被告知斑马是身上有斑纹的马,那么当孩子在动物园看到斑马时,就能顺利地识别出斑马. 零样本学习就是受人类这类学习方式的启发. 在零样本学习中,训练集(已见类别)和测试集(未见类别)样本类别是不相交的,通常通过一个语义空间,建立起已见类别和未见类别间的关系,进而实现知识的迁移. 其中语义空间通常由人工标注的属性、文本关键词或者词向量构成. 零样本图像分类就是在图像分类中应用了零样本学习的方法.

零样本图像分类方法通常分为零样本训练学习和零样本识别分类两个阶段. 现有零样本图像分类方法主要分为以下四类:(1)基于属性的学习[2-3]. 直接属性学习(Direct Attribute Prediction, DAP)[2]和间接属性学习(Indirect Attribute Prediction, IAP)[2]分别直接和间接学习单个语义属性的属性分类器,零样本识别阶段对未见类别样本预测该类别包含每个属性的概率,然后根据属性与类别的关系计算样本为各类别的分值,并将样本预测为得分最高的类别;(2)基于视觉空间到语义空间映射的学习[4-9]. 基于属性的标签嵌入方法(Attribute-based Label-Embedding, ALE)[4]在训练阶段学习一个兼容性函数,该函数用于衡量每一幅图像的视觉特征映射到语义空间后,和语义空间中每个类别语义属性向量之间的匹配度,确保每幅图像和所属类别语义属性向量的匹配度比其它类别的匹配度高. 测试阶段将兼容性得分最高的类别标签预测为该测试样本的标签. 结构化联合嵌入(Structured Joint Embedding, SJE)[5]受 ALE 的启发,使用了多种辅助语义信息源(包括传统属性、词向量、文本关键词等)替代人工标注属性. 与 ALE 类似,SJE 同样使用兼容性得分函数衡量视觉特征映射到语义空间后,与各类别语义表示的兼容性得分,不同的是语义空间包含多种信息源,因而需要学习多个映射;(3)基于语义空间到视觉空间映射的学习. Annadani 等人[10]提出将类别间的关系划分为语义相同的类别、语义相似的类别和语义不同的类别,在学习语义空间到视觉空间的映射时保留类别间的关系. 将视觉特征空间作为嵌入空间,一定程度上缓解了由少数枢纽点导致的枢纽点问题(Hubness problem). 枢纽点[11-13]是指这样一些点,它们是大多数其它点的最近邻点. 将视觉特征映射到语义空间,会产生一些枢纽点,在零样本识别阶段由于采用的是最邻近搜索方法,所以会降低识别的性能;(4)基于潜在子空间的学习[14-19]. 双视觉语义映射(Dual visual semantic Mapping Paths, DMaP)[14]学习视觉空间到语义空间的映射,同时抽取视觉空间中潜在类别级的流形构造新的语义空间,并结合原始的语义空间,不断迭代优化新的语义空间. 耦合字典学习(Coupled Dictionary Learning, CDL)[15]通过耦合字典学习框架分别学习两个字典,将视觉类别特征和语义类别特征分别映射到潜在子空间,并在潜在子空间学习过程中对齐语义类别特征和视觉类别特征.

虽然零样本图像分类的研究取得了一些进展,但是知识迁移过程中仍然存在以下主要问题:信息损失和域偏移问题. (1)信息损失问题,是指在学习图像的视觉特征与语义特征间映射的过程中,由于视觉特征和语义特征的维度相差较大,往往会出现一些具有判别能力的信息在知识迁移的过程中丢失的情况,从而影响最终的图像分类结果;(2)域偏移问题,是指由于零样本学习在训练过程中只用到了已见类别的信息,而训练类别和测试类别是不相交的,同时训练类别和测试类别往往差异可能很大,所以在测试的过程中会出现预测偏差导致域偏移问题.

针对信息损失和域偏移问题,本文提出了图像分类中基于子空间学习和重构的零样本分类(Zero-Shot Classification based on Subspace learning and Reconstruction, ZSCSR)方法. 该方法的主要贡献包含以下两个方面:

(1)首先假设语义空间和图像视觉空间具有相似的类别间关系,学习语义空间中已见类别和未见类别的关系,并将学习到的关系迁移到图像视觉空间,学习获得未见类别的视觉类别原型,缓解域偏移问题.

(2)基于子空间学习和重构的方法利用已见类别和未见类别的视觉特征和语义特征,学习一个共享的潜在子空间,该子空间既包含了图像视觉空间丰富的判别性信息也包含了语义空间中的类别间关系信息,同时在子空间学习过程中,通过重构,缓解知识迁移过程中信息损失的问题.

本文第 2 节给出问题定义和介绍相关工作;第 3 节提出基于子空间学习和重构的零样本分类方

法;第 4 节在几个通用零样本分类数据集上,通过实验对所提出的方法进行测试,并对实验结果和参数进行分析;最后对本文的工作进行总结.

2 问题定义和相关工作

本文的工作受到子空间学习和自编码器中重构思想的启发.下面分别给出本文的问题定义和相关工作简介.

2.1 问题定义

为了方便阐述,首先给出问题定义.设 $\boldsymbol{X}_s \in R^{d\times n_s}$ 为已见类样本视觉特征矩阵,其中 n_s 为已见类样本个数,d 为样本的特征维度.$\boldsymbol{S}_s \in R^{m\times c}$ 和 $\boldsymbol{S}_u \in R^{m\times t}$ 分别为已见类别的语义属性特征矩阵和未见类别的语义属性特征矩阵,其中 c 为已见类的类别个数,m 为语义特征的维度,t 为未见类的类别个数.$\boldsymbol{S}=[\boldsymbol{S}_s,\boldsymbol{S}_u]\in R^{m\times(c+t)}$ 是所有类别的语义特征矩阵.$Y_s=\{1,\cdots,c\}$ 为已见类的标签集,$Y_u=\{c+1,\cdots,c+t\}$ 为未见类的标签集,$Y_u \cap Y_s=\varnothing$.零样本图像分类就是给定 $\boldsymbol{X}_s$,$\boldsymbol{S}$,Y_s 和 Y_u 学习一个图像分类器 $f(\cdot)$,实现对未见样本 x_u 的分类,即 $f: x_u \rightarrow y_u$,其中 y_u 为未见样本 x_u 对应的标签.

2.2 子空间学习

在图像分类问题中,子空间学习是一种比较常见的方法.子空间学习通过学习一个合适的子空间,使得在原空间中不易识别或区分的图像,在子空间中类别差异扩大,或者子空间具备原始空间不具备的某些优势,进而将样本映射到子空间后获得更好的分类效果.由于零样本图像分类中的视觉特征通常是通过神经网络提取,而语义特征则是通过人工定义的属性或者文本中提取的关键词得到,所以视觉特征和语义特征的分布通常是不同的.如果直接学习获得视觉空间和语义空间之间的映射,通常知识迁移能力不强,导致零样本识别性能不好.通过子空间的学习,可以实现语义空间和视觉空间的对齐,获得较好的知识迁移能力.

双视觉语义映射(Dual visual-semantic Mapping paths,DMap)[14]通过学习一个语义子空间来对齐样本视觉空间和语义空间.DMap 首先学习获得样本视觉空间到原始语义空间映射,然后根据该映射,获得该类别样本视觉特征映射到语义空间后的均值,并与原始语义向量表示进行融合,迭代优化得到新的语义空间.潜在嵌入空间学习(Latent Embeddings,LatEm)[18]针对细粒度图像分类问题,提出将视觉和语义信息映射到一个多维向量空间,这个多维的向量空间即为学习的子空间.在子空间中将一些复杂的属性进行分解,来训练学习一组线性映射函数,不同的映射函数捕捉不同对象类的视觉特征,如颜色、形状或纹理等.针对不同的类别,LatEm 自动选择一组较好的线性函数模型来进行分类.耦合字典学习(Coupled Dictionary Learning,CDL)[15]采用字典学习方法获得一个子空间,在子空间中对齐视觉和语义结构.CDL 首先通过原型学习来学得已见类别在视觉空间的类别原型表示.然后通过字典学习,分别在视觉空间和语义空间中学习到一对字典的基,将视觉特征和语义特征映射到子空间,并约束同一类别的视觉特征和语义特征映射到子空间具有相同的特征表示,从而实现视觉空间和语义空间的结构对齐.

这些子空间学习虽然一定程度地提高了知识的迁移能力,但是在学习映射的过程中难免会出现信息损失的问题,尤其是一些子空间学习方法仅利用已见类别的视觉特征和语义特征来实现子空间学习.而丢失的信息可能对未见类别的识别有着重要的作用,因而影响迁移能力的提升.

2.3 自编码器

自编码器(Autoencoder,AE)是一个非监督学习算法,通常由三部分组成:编码器(encoder)、隐含层(hidden)和解码器(decoder).自编码器将输入表示 $\boldsymbol{X}$ 通过编码器编码到隐含层,再通过解码器解码回 $\boldsymbol{X}$.其中解码可以看作是重构的过程.自编码器的目标函数一般表示如式(1)所示:

$$\min_{\boldsymbol{W},\boldsymbol{W}^*} \|\boldsymbol{X}-\boldsymbol{W}^*\boldsymbol{W}\boldsymbol{X}\|_F^2 \tag{1}$$

其中,$\boldsymbol{X}\in R^{d\times n}$ 为输入样本,n 为输入样本个数,d 为样本特征维度.$\boldsymbol{W}\in R^{h\times d}$ 为编码矩阵,其中 h 为隐含层维度,$\boldsymbol{W}^*\in R^{d\times h}$ 为解码矩阵.

语义自编码器(Semantic Autoencode,SAE)[1]将自编码器的思想用于零样本图像分类,采用了一种简单的自编码器结构,只利用一层隐含层连接编码器和解码器.不同于一般的 AE,SAE 将隐含层定义为语义表示层,具有明确的语义.SAE 通过已见类别样本学习训练出编码和解码的映射矩阵,并利用该映射矩阵,将待识别的未见类别样本映射到语义空间进行识别.语义自编码器输入 $\boldsymbol{X}_s\in R^{d\times n_s}$ 是图像样本特征,通过映射矩阵 $\boldsymbol{W}\in R^{m\times d}$ 映射到隐含层为编码过程,然后再通过 $\boldsymbol{W}^{\mathrm{T}}\in R^{d\times m}$ 映射回视觉特征空间为解码过程,解码过程即为重构,使得重构

《计算机学报》

后的图像特征尽量地与原特征相近. SAE 目标函数如式(2)所示：

$$\min_{\boldsymbol{W}} \|\boldsymbol{X}_s-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{C}_s\|_F^2+\lambda\|\boldsymbol{W}\boldsymbol{X}_s-\boldsymbol{C}_s\|_F^2 \tag{2}$$

其中第一项和第二项分别对应解码过程和编码过程，$\boldsymbol{C}_s\in R^{m\times n_s}$ 是输入的已见样本语义属性矩阵，λ 是权重系数，调节第一项和第二项的重要性. 传统 AE 解码过程映射矩阵为 $\boldsymbol{W}^*$，而 SAE 中编码和解码过程是对称的，所以令 $\boldsymbol{W}^*=\boldsymbol{W}^{\mathrm{T}}$. 由于语义自编码器映射矩阵是由已见类别训练得出的，在测试阶段直接应用到未见类别，可能会由于已见类别和未见类别的分布不同，而导致学到的编码矩阵和解码矩阵泛化到未见类别的能力较弱.

3　基于子空间学习和重构的零样本分类方法

由于已见类别和未见类别是不相交的，其样本分布不同，如果在训练的过程中只利用已见类别的信息，那么学习到的模型往往不能较好地泛化到未见类别. 本文在训练过程中同时利用已见类别和未见类别信息，以提高模型的泛化性. 同时通过学习共享子空间，对齐视觉空间和语义空间，使得学习到的子空间中既包含语义空间中类别关系信息又包含视觉空间中可判别性信息，并且子空间学习过程中利用重构减少信息损失. 本文提出的基于子空间学习和重构的零样本分类方法(Zero-Shot Classification based on Subspace learning and Reconstruction，ZSCSR)同样包括训练学习阶段和零样本识别阶段. 训练学习阶段分为两步完成，训练学习框架如图 1 所示. 第一步学习语义空间中已见类别和未见类别的关系 $\boldsymbol{Q}$，并将该关系迁移到视觉空间，在视觉空间学习得到未见类的视觉类别原型. 第二步将学习到的未见类的视觉类别原型与已见类别的视觉类别原型融合得到所有类别在视觉空间的原型表示，然后采用融合重构的子空间学习，根据视觉空间和语义空间来学习共享子空间. 零样本识别阶段可以将未见类别样本分别映射到不同空间进行识别.

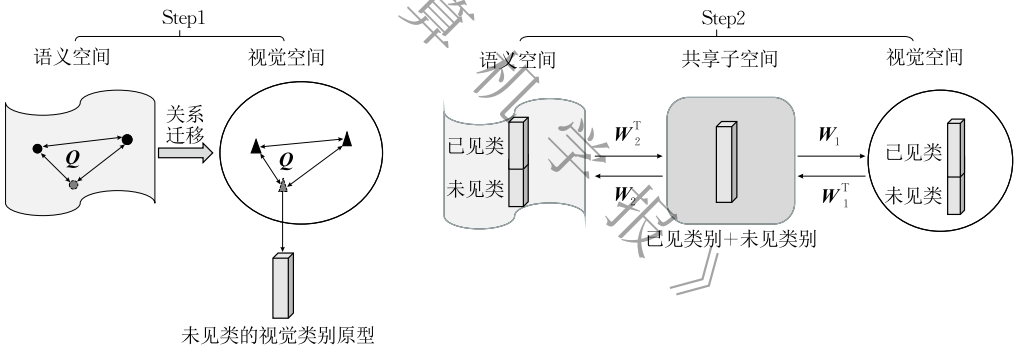

图 1　ZSCSR 的训练学习框架图

3.1　未见类别的类别原型学习

零样本分类在训练过程中，如果仅使用到已见类别的信息，学习到的模型不能很好地泛化到未见类别，从而导致分类准确率下降. 本文提出首先学习未见类别在视觉空间的类别原型，并将学习到的未见类别的视觉类别原型作为下一步的输入. 语义属性和视觉特征分别从不同的视角描述同一对象，因而语义空间和视觉空间上的类别关系是一致的. 虽然在语义空间和视觉空间中样本的特征维度不同，但是它们却有相似的类别间关系，即如果两个类别在语义空间中是相近的，那么它们在视觉空间中应该也是相近的. 由于零样本分类最终识别阶段是通过最近邻的方法实现未见类别样本识别，所以学习到接近于真实分布的类别原型，有助于提高识别准确率.

未见类别在视觉空间中的类别原型学习过程如图 2 所示. 图中语义空间为人工标注的属性向量，包含了相对全面的类别描述及类别间关系信息. 通过在语义空间中学习类别间关系矩阵 $\boldsymbol{Q}$，将其迁移到视觉空间，学习获得未见类别的视觉类别原型. 未见类别的类别原型学习的目标函数如式(3)所示：

$$\min_{\boldsymbol{Q},\boldsymbol{P}_u}\|\boldsymbol{S}_S\boldsymbol{Q}-\boldsymbol{S}_u\|_F^2+\|\boldsymbol{P}_S\boldsymbol{Q}-\boldsymbol{P}_u\|_F^2 \ \text{ s.t. } \|q^i\|_2^2\leq 1 \tag{3}$$

其中，$\boldsymbol{P}_S\in R^{d\times c}$ 是已见类别在视觉空间的类别原型矩阵，其中各类别原型取值为该类别下所有样本视觉特征向量的均值. $\boldsymbol{P}_u\in R^{d\times t}$ 是需要学习的未见类别在视觉空间的类别原型矩阵. $\boldsymbol{Q}\in R^{c\times t}$ 是学习到的语义空间中已见类与未见类的类别关系矩阵. $\boldsymbol{S}_S$ 是已见类别的类别语义特征矩阵，$\boldsymbol{S}_u$ 是未见类别的类别语义特征矩阵.

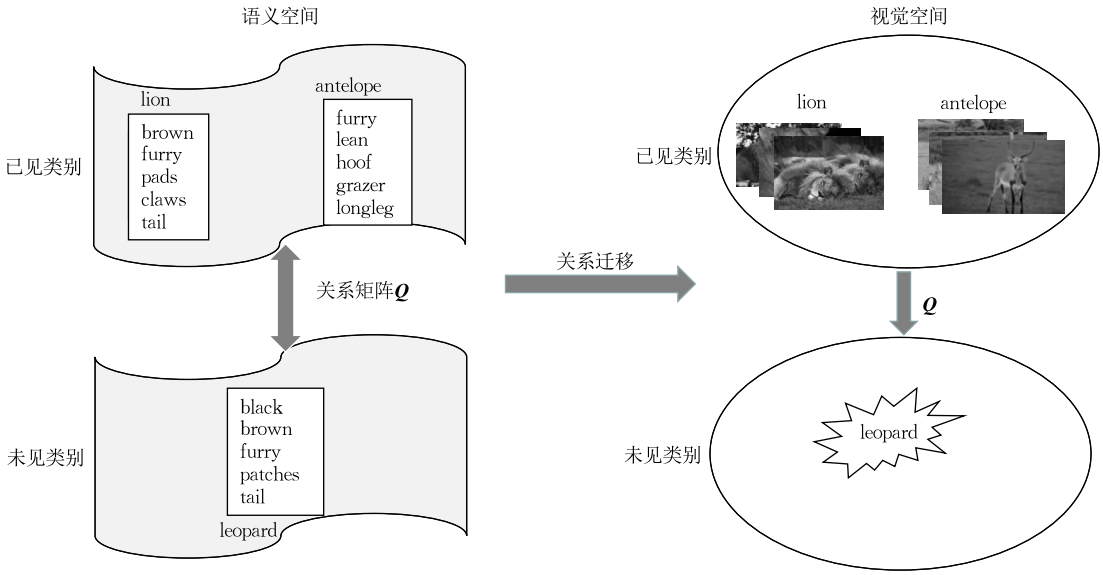

图 2 未见类别的视觉类别原型学习

本文采用交替优化方法求解目标函数(式(3)),即固定其它变量,求解某一变量,具体步骤如下:

(1)固定$\boldsymbol{P}_u$,更新$\boldsymbol{Q}$,得到优化函数如下式所示:

$$\min_{\boldsymbol{Q}}\|\boldsymbol{AQ}-\boldsymbol{B}\|_F^2+\alpha(\|q^i\|_2^2-1) \tag{4}$$

其中,$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{S}_S\\ \boldsymbol{P}_S\end{bmatrix}$,$\boldsymbol{B}=\begin{bmatrix}\boldsymbol{S}_u\\ \boldsymbol{P}_u\end{bmatrix}$.

直接对$\boldsymbol{Q}$求导可得

$$\boldsymbol{Q}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}+\alpha\boldsymbol{I})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B} \tag{5}$$

其中,$\boldsymbol{I}$为单位矩阵,α为拉格朗日乘子.

(2)固定$\boldsymbol{Q}$,更新$\boldsymbol{P}_u$,得到优化函数如下式所示:

$$\min_{\boldsymbol{P}_u}\|\boldsymbol{P}_s\boldsymbol{Q}-\boldsymbol{P}_u\|_F^2 \tag{6}$$

最后,根据$\boldsymbol{P}_u$,$\boldsymbol{P}_S$可得到$\boldsymbol{P}=[\boldsymbol{P}_s,\boldsymbol{P}_u]$.

3.2 融合重构的子空间学习

视觉空间是由图像自然的视觉特征构成的,包含了图像较为全面和细致的可判别信息.语义空间是由图像抽象的语义属性构成,包含了丰富的类别信息和类别关系信息.单纯的子空间学习在学习过程中,会造成部分信息丢失,而这些丢失的信息可能有助于未见类别样本的识别.本文提出的融合重构的子空间学习,利用所有类别(包括已见类别和未见类别)的语义和视觉信息,学习一个共享子空间,该子空间既具有视觉空间的可判别性信息,又具有语义空间的类别关系信息,同时利用重构,减少信息丢失.融合重构的子空间学习的学习框架如图 3 所示.

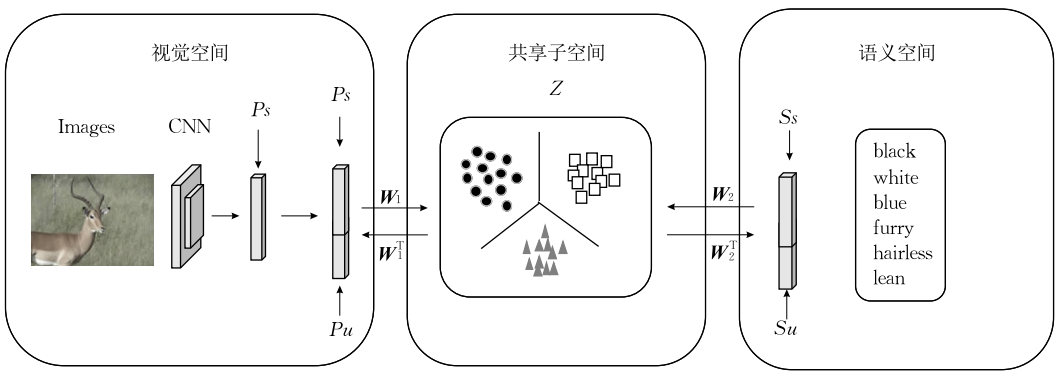

图 3 融合重构的子空间学习框架图

为了在子空间学习中对齐视觉空间和语义空间,令各类别的视觉特征和对应的语义特征映射到共享子空间中具有相同的特征表示.设要学习的共享子空间表示为$\boldsymbol{Z}$,学习视觉空间到共享子空间映射的目标函数如式(7)所示:

$$\min_{\boldsymbol{W}_1,\boldsymbol{Z}}\lambda_1\|\boldsymbol{W}_1\boldsymbol{P}-\boldsymbol{Z}\|_F^2+\|\boldsymbol{P}-\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{Z}\|_F^2 \tag{7}$$

其中,$\boldsymbol{Z}\in R^{k\times(c+t)}$是包含已见类与未见类的所有类别在共享子空间中的表示,k为在共享子空间中的维度,c和t分别是已见类和未见类的类别个数,$\boldsymbol{P}=[\boldsymbol{P}_s,\boldsymbol{P}_u]\in R^{d\times(c+t)}$为包含已见类和未见类的所

有类别的视觉类别原型矩阵. 由于训练过程中无法得到未见类别的视觉特征，所以 $\boldsymbol{P}_u$ 由 3.1 节学习获得. $\boldsymbol{W}_1 \in R^{k\times d}$ 是视觉空间到共享子空间的映射矩阵. 式(7)中第一项为视觉空间到共享子空间的映射，第二项为根据共享子空间重构视觉空间，权重系数 λ_1 调节这两项之间的比重. 通过空间映射差异最小化，实现子空间学习获得视觉空间判别性信息，同时通过重构误差最小化，实现子空间尽可能多地保存原始信息，缓解了知识迁移过程中的信息损失问题. 由于映射和重构分别使用了矩阵 $\boldsymbol{W}_1$ 和它的转置矩阵 $\boldsymbol{W}_1^{\mathrm{T}}$，因而该目标函数隐式地约束了 $\boldsymbol{W}_1$ 不会太大，而是在一个合理的范围内.

类似地，学习语义空间到共享子空间映射的目标函数如式(8)所示：

$$\min_{\boldsymbol{W}_2,\boldsymbol{Z}} \lambda_2 \left\|\boldsymbol{W}_2\boldsymbol{S}-\boldsymbol{Z}\right\|_F^2+\left\|\boldsymbol{S}-\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{Z}\right\|_F^2 \tag{8}$$

同样 $\boldsymbol{S}=[\boldsymbol{S}_s,\boldsymbol{S}_u]\in R^{m\times(c+t)}$ 为包含已见类和未见类的所有类别的语义特征矩阵. $\boldsymbol{W}_2\in R^{k\times m}$ 是语义空间到共享子空间的映射矩阵. 式(8)中第一项为语义空间到共享子空间的映射，第二项为根据共享子空间重构语义空间，权重系数 λ_2 调节这两项之间的比重. 通过空间映射差异最小化，实现子空间学习获得语义空间类别间关系信息，同时通过重构误差最小化，实现子空间尽可能多地保存原始信息，缓解了知识迁移过程中的信息损失问题.

综上，融合重构的子空间学习的目标函数如式(9)所示：

$$\begin{aligned}\min_{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{Z}} \lambda_1 \left\|\boldsymbol{W}_1\boldsymbol{P}-\boldsymbol{Z}\right\|_F^2+\left\|\boldsymbol{P}-\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{Z}\right\|_F^2+\\ \lambda_2 \left\|\boldsymbol{W}_2\boldsymbol{S}-\boldsymbol{Z}\right\|_F^2+\left\|\boldsymbol{S}-\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{Z}\right\|_F^2\end{aligned} \tag{9}$$

本文采用交替迭代法求解目标函数(式(9)). 具体求解过程如下：

(1) 固定 $\boldsymbol{W}_2$，$\boldsymbol{Z}$，更新 $\boldsymbol{W}_1$，得到优化函数如下式所示：

$$\min_{\boldsymbol{W}_1} \lambda_1 \left\|\boldsymbol{W}_1\boldsymbol{P}-\boldsymbol{Z}\right\|_F^2+\left\|\boldsymbol{P}-\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{Z}\right\|_F^2 \tag{10}$$

可直接对 $\boldsymbol{W}_1$ 求导，令 $\boldsymbol{A}_1=\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$，$\boldsymbol{B}_1=\lambda_1\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}$，$\boldsymbol{C}_1=(\lambda_1+1)\boldsymbol{Z}\boldsymbol{P}^{\mathrm{T}}$ 得：

$$\boldsymbol{A}_1\boldsymbol{W}_1+\boldsymbol{W}_1\boldsymbol{B}_1=\boldsymbol{C}_1 \tag{11}$$

上式可直接由 Sylvester 方程求解.

(2) 固定 $\boldsymbol{W}_1$，$\boldsymbol{Z}$，更新 $\boldsymbol{W}_2$，得到优化函数如下式所示：

$$\min_{\boldsymbol{W}_2} \lambda_2 \left\|\boldsymbol{W}_2\boldsymbol{S}-\boldsymbol{Z}\right\|_F^2+\left\|\boldsymbol{S}-\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{Z}\right\|_F^2 \tag{12}$$

直接对 $\boldsymbol{W}_2$ 求导，令 $\boldsymbol{B}_2=\lambda_2\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}}$，$\boldsymbol{C}_2=(\lambda_2+1)\boldsymbol{Z}\boldsymbol{S}^{\mathrm{T}}$ 得：

$$\boldsymbol{A}_1\boldsymbol{W}_2+\boldsymbol{W}_2\boldsymbol{B}_2=\boldsymbol{C}_2 \tag{13}$$

上式可直接由 Sylvester 方程求解.

(3) 固定 $\boldsymbol{W}_1$，$\boldsymbol{W}_2$，更新 $\boldsymbol{Z}$，得到优化函数如下式所示：

$$\begin{aligned}\min_{\boldsymbol{Z}} \lambda_1 \left\|\boldsymbol{W}_1\boldsymbol{P}-\boldsymbol{Z}\right\|_F^2+\left\|\boldsymbol{P}-\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{Z}\right\|_F^2+\\ \lambda_2 \left\|\boldsymbol{W}_2\boldsymbol{S}-\boldsymbol{Z}\right\|_F^2+\left\|\boldsymbol{S}-\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{Z}\right\|_F^2\end{aligned} \tag{14}$$

得到 $\min_{\boldsymbol{Z}} \left\|\boldsymbol{A}_2-\boldsymbol{B}_3\boldsymbol{Z}\right\|_F^2$，

其中，$\boldsymbol{A}_2=\begin{bmatrix}\lambda_1\boldsymbol{W}_1\boldsymbol{P}\\ \lambda_2\boldsymbol{W}_2\boldsymbol{S}\\ \boldsymbol{P}\\ \boldsymbol{S}\end{bmatrix}$，$\boldsymbol{B}_3=\begin{bmatrix}\lambda_1\boldsymbol{I}\\ \lambda_2\boldsymbol{I}\\ \boldsymbol{W}_1^{\mathrm{T}}\\ \boldsymbol{W}_2^{\mathrm{T}}\end{bmatrix}$，$\boldsymbol{I}$ 为单位矩阵.

对 $\boldsymbol{Z}$ 直接求导得

$$\boldsymbol{Z}=(\boldsymbol{B}_3^{\mathrm{T}}\boldsymbol{B}_3)^{-1}\boldsymbol{B}_3^{\mathrm{T}}\boldsymbol{A}_2 \tag{15}$$

3.3　算法流程

本文基于子空间学习和重构的零样本分类方法中训练学习阶段流程如算法 1 所示.

算法 1.　基于子空间学习和重构的零样本学习.

输入：已见类别的类别语义矩阵 $\boldsymbol{S}_s$，已见类别所有样本在视觉特征空间的特征矩阵 $\boldsymbol{X}_s$，未见类别的类别语义矩阵 $\boldsymbol{S}_u$，最大迭代次数 I

输出：$\boldsymbol{W}_1$，$\boldsymbol{W}_2$，$\boldsymbol{Z}$

过程：优化更新过程

1. 初始化 $\boldsymbol{Q}$，$\boldsymbol{P}_u$，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$，$\boldsymbol{Z}$

Step1.

2. 通过 $\boldsymbol{X}_s$ 计算已见类别在视觉特征空间的类别原型 $\boldsymbol{P}_s$，各维度的值是该类别在视觉特征空间中的所有样本在该维度上的均值
3. 根据式(5)计算类别间关系矩阵 $\boldsymbol{Q}$
4. 根据式(6)计算未见类别的类别原型 $\boldsymbol{P}_u$

Step2.

5. 计算所有类别的类别语义矩阵 $\boldsymbol{S}\in R^{m\times(c+t)}$ 和所有类别的类别原型矩阵 $\boldsymbol{P}\in R^{d\times(c+t)}$
6. FOR $i=1$ to I
7. 　根据式(11)计算映射矩阵 $\boldsymbol{W}_1$
8. 　根据式(13)计算映射矩阵 $\boldsymbol{W}_2$
9. 　根据式(15)计算所有类别在潜在共享子空间中的表示矩阵 $\boldsymbol{Z}$
10. END FOR

3.4　零样本识别

本文在零样本识别阶段采用最近邻方法识别未见类别样本. 由于本文方法涉及共享子空间、视觉空间和语义空间. 下面给出将待识别未见类别样本 x_u 映射到不同空间下进行识别的具体步骤.

3.4.1　共享子空间中识别未见类别

首先将 x_u 从视觉空间映射到共享子空间中，即 $\hat{z}_u=\boldsymbol{W}_1 x_u$.

然后将语义空间中的类别语义矩阵 $\boldsymbol{S}_u$ 映射到

共享子空间中，得到未见类别在共享子空间中的新的表示 $\boldsymbol{Z}_u=\boldsymbol{W}_2\boldsymbol{S}_u$.

最后，预测 x_u 对应的类别标签 y 为

$$y=\arg\min_j D(\hat{z}_u,\boldsymbol{z}_u^j),$$

其中，D 是一个距离函数，本文采用的是余弦距离. $\boldsymbol{z}_u^j$ 是矩阵 $\boldsymbol{Z}_u$ 中的第 j 列向量，即第 j 个类别在共享子空间中的特征表示.

3.4.2 视觉空间识别未见类别

在视觉空间的识别未见类别样本 x_u 有以下两种方法：

(1) 方法一

利用 3.2 节学习到的映射矩阵将语义表示映射到视觉空间来进行未见类别的识别.

首先将未见类别的语义表示矩阵映射到视觉空间得到预测的各未见类别的视觉特征矩阵 $\hat{\boldsymbol{P}}_u=\boldsymbol{W}_1^{\mathrm{T}}(\boldsymbol{W}_2\boldsymbol{S}_u)$.

然后预测 x_u 对应的类别标签 $y=\arg\min_j D(x_u,\hat{\boldsymbol{p}}_u^j)$，其中 $\hat{\boldsymbol{p}}_u^j$ 是矩阵 $\hat{\boldsymbol{P}}_u$ 中的第 j 列向量，即第 j 个未见类别的视觉类别原型向量.

(2) 方法二

直接根据 3.1 节学习到的未见类别的视觉类别原型 $\boldsymbol{P}_u$，预测未见类别样本 x_u 对应的类别标签 $y=\arg\min_j D(x_u,\boldsymbol{p}_u^j)$，其中 $\boldsymbol{p}_u^j$ 是矩阵 $\boldsymbol{P}_u$ 中的第 j 列向量，即第 j 个未见类别的视觉类别原型向量.

3.4.3 语义空间中识别未见类别

首先将 x_u 从视觉空间映射到语义空间中，即 $\hat{s}_u=\boldsymbol{W}_2^{\mathrm{T}}(\boldsymbol{W}_1 x_u)$.

然后预测 x_u 对应的类别标签 $y=\arg\min_j D(\hat{s}_u,\boldsymbol{s}_u^j)$，其中 $\boldsymbol{s}_u^j$ 为矩阵 $\boldsymbol{S}_u$ 中的第 j 列向量，即第 j 个未见类别的语义表示向量.

4 实验结果与分析

4.1 数据集介绍及实验设置

本文实验数据集采用零样本图像分类中普遍采用的四个公共基准数据集：Animals with Attributes 2 (简记为 AwA2)[20]，CUB-200-2011 Birds (简记为 CUB)[21]，aPascal&aYahoo (简记为 aP&Y)[22] 和 SUN Attribute (简记为 SUN)[23]. 数据集 AwA2 的属性维度为 85 维，其中 40 个已见类别的 30337 张图片作为训练集和 10 个未见类别的 6985 张图片作为测试集，共计 37322 个图像样本. 数据集 CUB 的属性维度为 312 维，其中 150 个已见类别的 8855 张图片作为训练集和 50 个未见类别的 2933 张图片作为测试集，共计 11788 个图像样本. 数据集 SUN 的属性维度为 102 维，其中 645 个已见类别的 12900 张图片作为训练集和 72 个未见类别的 1440 张图片作为测试集，共计 14340 个图像样本. 数据集 aP&Y 的属性维度为 64 维，其中 20 个已见类别的 12695 张图片作为训练集和 12 个未见类别的 2644 张图片作为测试集，共计 15339 个图像样本.

实验设置：本文所有样本的视觉特征均采用 GoogleNet 提取的 1024 维的特征. 语义特征均采用各数据集中存储的人工定义的属性特征. 在实验过程中设置共享子空间维度为所有类别的总个数，首先初始化所有类别在共享子空间中的表示 $\boldsymbol{Z}$ 为所有类别的相似度矩阵.

4.2 主流方法的对比实验

为了验证本文所提的 ZSCSR 方法的有效性，本文分别和相关的 10 种主流方法进行了对比实验：Direct Attribute Prediction (DAP)[2]，Embarrassingly Simple Zero Shot Learning (ESZSL)[24]，Attribute Label Embedding (ALE)[4]，Structured Joint Embedding (SJE)[5]，Latent Attribute Dictionary (LDA)[17]，Predicting Visual Exemplars (EXEM)[25]，Semantic Autoencoder (SAE)[1]，Semantics-Preserving Adversarial Embedding Network (SP-AEN)[7]，Synthesized Classifiers (SYNC)[26] 和 Coupled Dictionary Learning (CDL)[15]. 同时为了验证本文提出框架每一部分的有效性，本文对不同的子任务进行实验对比，ZSCSR-E 是指删去原始空间到子空间的映射部分，ZSCSR-D 是指删去子空间重构原始空间部分，ZSCSR-P 是指删去学习未见类别视觉特征原型的部分. 表 1 为对比实验结果.

表 1 不同方法的分类准确率 (单位：%)

Method	AwA2	CUB	aP&Y	SUN
DAP	50.5	37.4	20.1	39.9
ESZSL	75.3	31.4	24.2	49.8
ALE	62.5	43.9	40.0	55.3
SJE	73.9	47.6	36.7	40.5
LDA	77.1	45.9	14.1	55.8
EXEM	70.5	46.2	42.3	60.0
SAE	79.6	49.1	34.2	54.9
SP-AEN	80.3	46.6	24.1	59.2
SYNC	64.0	48.7	23.6	53.3
CDL	65.6	50.2	11.5	43.1
ZSCSR	**84.1**	**52.7**	**51.6**	**63.0**
ZSCSR-E	43.3	43.2	40.7	49.2
ZSCSR-D	77.7	46.7	46.7	57.4
ZSCSR-P	80.0	46.9	42.5	60.3

《计算机学报》

表 1 中，DAP 是零样本图像分类中较为经典的方法，训练过程中针对每个属性训练对应的属性分类器，对测试样本直接预测各属性的概率. 但是属性分类器是分开训练的，并没有学习到属性间的关系，ZSCSR 相对 DAP 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 33.6%，15.3%，31.5%和 23.1%. ESZSL 引入了一个双层的线性模型，分别建模特征与语义之间的关系和语义与标签之间的关系，在双层的线性模型知识迁移过程中会存在信息损失的问题. 相对 ESZSL，ZSCSR 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 8.8%，21.3%，27.4% 和 13.2%. ALE 和 SJE 都是通过学习兼容性函数，来度量图像和语义空间的兼容性. SJE 是在 ALE 的基础上，联合学习多个兼容性函数来帮助预测未见类别的标签. ZSCSR 相对于 ALE 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 21.6%，8.8%，11.6%和 7.7%. ZSCSR 相对于 SJE 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 10.2%，5.1%，14.9%和 22.5%. LDA 将学习获得的潜在属性空间作为语义空间，潜在属性为已见属性的线性组合，由于该模型通过已见类别进行训练，在预测未见类别时，使得未见类别易偏向于已见类，使分类产生错误. ZSCSR 相对于 LDA 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 7%，6.8%，37.5%和 7.2%. EXEM 将语义属性映射到视觉空间，使用核回归的方法来匹配语义属性对应的视觉特征聚类中心. ZSCSR 相对于 EXEM 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 13.6%，6.5%，9.3%和 3%. SAE 和 SP-AEN 都采用了重构的思想，用编码和解码两个过程来学习视觉特征空间到语义空间的映射. SP-AEN 利用已见类别样本作为训练集，将子空间学习分成了两个子任务，分别实现重构和分类，并且在两个子任务中利用对抗学习实现零样本分类. ZSCSR 相对于 SAE 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 4.5%，3.6%，17.4%和 8.1%，ZSCSR 相对于 SP-AEN 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 3.8%，6.1%，27.5%和 3.8%. SYNC 和 CDL 都是采用了字典学习的方法. SYNC 通过对这些字典的基的组合来合成未见类别分类器. CDL 则是通过字典的基来学习子空间，在子空间中对齐语义和特征信息. ZSCSR 相对于 CDL 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 18.5%，2.5%，40.1%和 19.9%. ZSCSR 相对于 SYNC 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 20.1%，4%，28%和 9.7%.

相对于 ZSCSR-E，ZSCSR 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 40.8%，9.5%，10.9%和 13.8%. 相对于 ZSCSR-D，ZSCSR 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 6.4%，6%，4.9%和 5.6%. 相对于 ZSCSR-P，ZSCSR 在数据集 AwA2、CUB、aP&Y 和 SUN 上的分类准确率分别提高了 4.1%，5.8%，9.1%和 2.7%. 从实验结果看，ZSCSR 同时利用已见类别和未见类别信息，通过学习视觉空间和语义空间到子空间的映射，学习到了原始空间中的可判别信息和类别关系信息，并且利用重构，减少信息损失，在未见样本识别上获得较好的识别效果.

4.3　不同的空间中零样本识别对比实验

本文提出的方法在识别阶段，可以分别在三个空间下进行零样本识别. 为了验证不同空间下零样本识别的效果，本文进行了对比实验，对比实验的方法分别为 3.4 节中介绍的 4 种方法：共享子空间中的识别方法，语义空间中的识别方法，视觉空间下的识别方法一和视觉空间下的识别方法二. 实验结果如表 2 所示.

表 2　不同空间中识别方法的识别准确率（单位：%）

	AwA2	CUB	SUN	aP&Y
语义空间	79.7	45.5	54.8	20.4
共享子空间	80.6	**52.7**	61.7	**51.6**
视觉空间方法一	**84.1**	51.3	**63.0**	45.3
视觉空间方法二	71.6	35.4	58.7	15.4

由表 2 可以看出，对于不同的数据集，每个空间的表示能力不同. 实验结果显示，在 4 个数据集上，视觉空间中方法一和共享子空间的识别性能均高于语义空间的，说明视觉空间比语义空间包含更多的判别性信息，共享子空间通过学习了视觉空间和语义空间的互补信息，相较于语义空间有了一定程度的提升. 对于数据集 CUB 和 aP&Y，在共享子空间中的识别性能更高，因为在共享子空间中同时学习到了图像视觉空间的判别性信息，也学习到了语义空间类别间关系的信息，两个空间中的信息融合互补，所以在共享子空间中能够更好地对未见类别进行识别. 对于数据集 AwA2 和 SUN，在视觉空间中方法一的识别能力高于共享子空间. 在数据集

AwA2 中，每个类别包含了足够多的图片数量，具有足够的判别性信息能够很好地概括每个类别，但是类别数较少，没有足够的类别间关系信息，所以图像视觉空间的识别能力更好. 数据集 SUN 虽然包含了较多的类别数，但是由于 SUN 是场景数据集，涵盖的类别范围较广，用于描述每个类别的语义属性维度却没有足够多，所以 SUN 的语义并没有足够好地描述所有类别，因而语义空间中的关系信息对分类结果的提升影响不大.

视觉空间方法二是仅利用 3.1 节方法学习未见类别的视觉类别原型，并未进一步利用 3.2 节融合重构的子空间学习，识别性能虽然相较于其它方法来说差一点，但是也具有一定的识别能力，说明本文通过关系矩阵学习到的未见类别的类别原型能够较好地拟合未见类别的真实分布.

为了进一步验证本文方法的适应性，本文将预测标签集扩展到包括已见类和未见类的所有类别的标签集，分别在四个不同的数据集上进行了对比实验，实验结果如表 3 所示.

表 3 预测标签集扩展后的不同空间中识别方法的识别准确率 （单位：%）

	AwA2	CUB	SUN	aP&Y
语义空间	47.6	21.1	38.4	15.6
共享子空间	50.3	32.3	40.0	**38.2**
视觉空间方法一	**56.4**	**33.2**	**49.6**	30.2
视觉空间方法二	42.7	20.4	32.5	10.9

对比表 2 和表 3，可以看出，当标签集扩展到包含了所有已见类别和未见类别的标签集时，分类的准确率会有所下降，这是因为部分未见类别图像会被识别成与之相似的已见类别. AWA2 数据集在各空间中识别准确率下降最多，其次是 CUB 数据集. 因为 AWA2 数据集为动物数据集，CUB 为鸟类数据集，有较多的相似类别，所以会使得部分未见类别错误分类成相似的已见类别.

4.4 参数分析

为检验 ZSCSR 中各参数（λ_1，λ_2 和迭代次数 i）对模型性能的影响，本节在 SUN 数据集上进行对比实验. 实验中分别固定其它参数，调节其中一个参数，实验分别给出在不同空间中零样本识别的准确率. 以下实验结果中视觉空间的识别方法均为视觉空间方法一的结果.

首先对参数 λ_1 对模型性能的影响进行对比实验，设参数 $\lambda_2=100$，迭代次数 $i=31$. 实验结果如表 4 所示，可以看出 λ_1 对三个空间中的识别效果的影响是不同的. λ_1 是调节视觉空间到共享子空间映射和重构过程的重要性参数. 当 λ_1 等于 0 时，相当于在视觉空间到共享子空间学习映射的过程中，没有编码过程只有解码过程，所以在学习共享子空间表示 $\boldsymbol{Z}$ 的过程中，没有学习到视觉特征空间中类别原型的判别性信息. λ_1 等于 1 时，编码和解码过程重要性相同，此时视觉空间获得了最高的识别准确率，随着 λ_1 的增大，当 λ_1 等于 10 时，语义空间获得了最高的识别准确率. 当 λ_1 等于 100 时共享子空间获得了最高的识别准确率，说明此时在共享子空间中很好地学习到了视觉特征空间中的判别性信息，同时通过重构的过程减少了视觉特征空间到共享子空间映射过程中的信息损失. 当 $\lambda_1\to\infty$ 时，几乎忽略了重构过程，使得在映射的过程中信息损失增大，最终识别准确率降低.

表 4 不同 λ_1 下在三个不同空间中的识别准确率 （单位：%）

λ_1	0	1	10	100	200	300	400	500	600	700	800	900	1000	2000
共享子空间	50.7	53.1	55.0	**61.7**	58.9	58.3	58.7	59.0	59.2	59.2	59.5	59.2	58.7	51.2
视觉空间	57.1	**63.0**	61.8	57.8	56.8	57.4	58.4	59.2	59.7	59.7	59.9	60.0	60.2	57.6
语义空间	45.8	53.5	**54.8**	50.0	45.6	43.2	43.1	42.9	42.9	43.1	42.8	42.9	42.8	40.0

然后对参数 λ_2 对模型性能的影响进行对比实验，设参数 $\lambda_1=1$，迭代次数 $i=31$，实验结果如表 5 所示. 可以看到，共享子空间在 λ_2 等于 10 时，获得了最高的识别准确率. 在视觉空间和语义空间中，当 λ_2 等于 100 时，均获得了最高的识别准确率. λ_2 是调节语义空间到共享子空间映射的编码过程和重构的解码过程的重要性参数. λ_2 等于 0 时，仅有重构过程，无法学习到原语义空间中的类别信息. 随着 λ_2 的增大，编码过程的重要性逐渐增强，能够通过编码过程学习语义空间的信息，同时解码的重构过程减少了语义空间到共享子空间映射过程中的信息损失，更好地学习了原语义空间的类别信息以及类别间关系信息. 当 λ_2 继续增大时，重构过程的相对重要性降低，直至被忽略，使得原语义空间信息损失增加，导致识别准确率下降.

表 5　不同 λ_2 下三个不同空间中的识别准确率　　(单位：%)

λ_2	0	1	10	100	200	300	400	500	600	700	800	900	1000	2000
共享子空间	21.7	47.9	**61.7**	57.5	53.1	53.3	53.3	53.1	52.8	52.4	52.3	52.2	52.2	51.6
视觉空间	15.8	46.0	61.8	**63.0**	62.4	62.2	62.2	62.1	62.1	62.2	61.9	61.8	61.9	61.7
语义空间	12.4	39.5	53.5	**54.8**	53.4	53.1	52.8	52.5	52.2	51.9	52.0	52.0	52.1	50.8

最后对迭代次数 i 对模型性能的影响进行对比实验，设参数 $\lambda_1=1$，$\lambda_2=100$. 实验结果如表 6 所示，可以看出共享子空间中，识别准确率在迭代 28 次左右时达到收敛；在视觉空间中，识别准确率在迭代 10 次左右时达到收敛；在语义空间中，识别准确率在迭代 31 次左右时达到收敛.

表 6　不同迭代次数 i 下的识别准确率　　(单位：%)

i	1	4	7	10	13	16	19	22	25	28	31	34
共享子空间	35.0	51.2	53.8	57.6	58.8	59.3	60.7	61.1	61.4	**61.7**	61.7	61.7
视觉空间	44.9	62.0	62.3	**63.0**	63.0	63.0	63.0	63.0	62.9	62.9	63.0	63.0
语义空间	28.0	50.5	51.4	52.1	52.7	53.0	53.5	53.9	54.2	54.5	**54.8**	54.8

5　总　结

随着大规模图像数据集(例如 ImageNet)的出现，图像分类研究取得了极大的进展. 然而，新的图像类别和新的分类需求(例如细粒度图像分类)不断涌现. 获取足够的新类别的标注样本成本太大，甚至在特定领域是非常困难的. 零样本图像分类具有非常重要的研究价值.

针对零样本分类中知识迁移过程中信息损失和域偏移问题，本文充分利用已见类别和未见类别信息，学习语义空间的类别间关系，同时将学习到的类别间关系迁移到视觉空间，从而学习获得未见类别的视觉类别原型. 同时通过共享子空间的学习和重构的思想，学习获得共享子空间，在共享子空间中保存语义空间的关系信息和视觉空间中的判别性信息，两个空间中的信息互补，从而能够更好地表示各类别，提升零样本识别效果.

参　考　文　献

[1] Kodirov E, Xiang Tao, Gong Shaogang. Semantic autoencoder for zero-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 4447-4456

[2] Lampert C H, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Florida, USA, 2009: 951-958

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

[4] Akata Z, Perronnin F, Harchaoui Z. Label-embedding for attribute-based classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 819-826

[5] Akata Z, Reed S, Walter D, et al. Evaluation of output embeddings for fine-grained image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2927-2936

[6] Morgado P, Vasconcelos N. Semantically consistent regularization for zero-shot recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 2037-2046

[7] Chen Long, Zhang Hanwang, Xiao Jun, et al. Zero-shot visual recognition using semantics-preserving adversarial embedding networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah, USA, 2018: 1043-1052

[8] Song Jie, Shen Chengchao, Yang Yezhou, et al. Transductive unbiased embedding for zero-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah, USA, 2018: 1024-1033

[9] Akata Z, Perronnin F, Harchaoui Z, Schmid C. Label-embedding for image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(7): 1425-1438

[10] Annadani Y, Biswas S. Preserving semantic relations for zero-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Utah, USA, 2018: 7603-7612

[11] Lazaridou A, Dinu G, Baroni M. Hubness and pollution: Delving into cross-space mapping for zero-shot learning//Proceedings of the Meeting of the Association for Computational Linguistics & the International Joint Conference on Natural Language Processing. Beijing, China, 2015: 270-280

[12] Dinu G, Lazaridou A, Baroni M. Improving zero-shot learning

《计算机学报》

by mitigating the hubness problem//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 10-20

[13] Shigeto Y, Suzuki I, Hara K, et al. Ridge regression, hubness, and zero-shot learning//Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Porto, Portugal, 2015: 135-151

[14] Li Yanan, Wang Donghui, Hu Huanhang, et al. Zero-shot recognition using dual visual semantic mapping paths//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 5207-5215

[15] Jiang Huajie, Wang Ruiping, Shan Shiguang, Chen Xilin. Learning class prototypes via structure alignment for zero-shot recognition//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 121-138

[16] Fu Yanwei, Hospedales T M, Xiang Tao, Gong Shaogang. Transductive multi-view zero-shot learning. Pattern Analysis and Machine Intelligence, 2015, 37(11): 2332-2345

[17] Jiang Huajie, Wang Ruiping, Shan Shiguang, et al. Learning discriminative latent attributes for zero-shot classification//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 4233-4242

[18] Xian Yongqin, Akata Z, Sharma G, et al. Latent embeddings for zero-shot classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 69-77

[19] Fu Yanwei, Hospedales T M, Xiang Tao, et al. Transductive multi-view embedding for zero-shot recognition and annotation //Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 584-599

[20] Xian Yongqin, Lampert C H, Schiele B, Akata Z. Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 40(9): 2251-2265

[21] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 dataset. Pasadena, USA: California Institute of Technology Computation & Neural Systems, Technical Report CNS-TR-2011-001, 2011

[22] Farhadi A, Endres I, Hoiem D, Forsyth D. Describing objects by their attributes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Florida, USA, 2009: 1778-1785

[23] Patterson G, Xu Chen, Su Hang, Hays J. The SUN attribute database: Beyond categories for deeper scene understanding. International Journal of Computer Vision, 2014, 108(1-2): 59-81

[24] Romera-Paredes B, Torr P H S. An embarrassingly simple approach to zero-shot learning//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 2152-2161

[25] Changpinyo S, Chao Wei-Lun, Sha Fei. Predicting visual exemplars of unseen classes for zero-shot learning//Proceedings of the International Conference on Computer Vision. Venice, Italy, 2017: 3496-3505

[26] Changpinyo S, Chao Wei-Lun, Gong Boqing, Sha Fei. Synthesized classifiers for zero-shot learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5327-5336

《计算机学报》

ZHAO Peng, Ph. D., associate professor, M. S. supervisor. Her research interests include machine learning and image understanding.

WANG Chun-Yan, M. S. candidate. Her research interests include machine learning and image classification.

ZHANG Si-Ying, M. S. candidate. Her research interests include machine learning and image classification.

LIU Zheng-Yi, Ph. D., associate professor, M. S. supervisor. Her research interests include machine learning and computer vision.

Background

Image classification is a very important task in computer vision and image understanding. Traditional image classifier can only classify the samples from the seen categories which have appeared in the training dataset. But in real-world applications, new categories continue to emerge. It is very time-consuming to collect enough labeled samples of the new category and retrain the classifier. As we know, humans are very good at recognizing objects without seeing any visual sample. Inspired by the above ability of human, zero-shot classification emerges and has become a very important topic in recent years.

Zero-shot classification has shown to be of utility in various applications, such as face recognition, action recognition, activity recognition, object recognition, event detection, and

so on. Zero-shot classification aims to utilize the semantic prototypes of all categories and the visual feature of the data from the seen categories to classify the data from the unseen categories. The seen categories refer to the categories with sufficient labeled data. The unseen categories refer to the new categories without labeled data. The semantic prototype means the embedded label representation in a semantic space. Such a semantic space can be a semantic attribute space or a semantic word vector space. Zero-shot classification can be taken as a special case of transfer learning, where the seen categories are the source domain categories and the unseen categories are the target domain categories. The key problems in zero-shot classification are what are the relationship between the seen categories and the unseen categories and how to classify the unseen data accurately. Most existing zero-shot classification methods learn a mapping function from the visual space to the semantic embedding space only using the visual features of the labeled training data from the seen categories. There are two main problems in the zero-shot classification: domain shift and information loss. In this paper, we present a novel zero-shot classification approach based on subspace learning and reconstruction for image classification (Zero-Shot Classification based on Subspace learning and Reconstruction, ZSCSR), which relieves the problems of domain shift and information loss in the transfer learning of zero-shot classification.

This paper is supported by the National Natural Science Foundation of China (Grant No. 61602004), the Natural Science Foundation of the Education Department of Anhui Province (Grant Nos. KJ2018A0013, KJ2017A011), the Natural Science Foundation of Anhui Province (Grant Nos. 1908085MF188, 1908085MF182), and the Key Research and Development Program of Anhui Province (Grant No. 1804d08020309).

《计算机学报》