基于低秩堆栈式语义自编码器的零样本学习*

冉瑞生†，董殊宏，李 进，王 宁
（重庆师范大学 计算机与信息科学学院，重庆 401331）

摘 要： 在图像分类领域，现有的深度学习等方法在训练时需要大量有标注的数据样本，且无法识别在训练阶段未出现的类别。零样本学习能有效缓解此类问题。本研究基于堆栈式自编码器和低秩嵌入，提出了一种新的零样本学习方法，即基于低秩嵌入的堆栈语义自编码器（low-rank stacked semantic auto-encoder，LSSAE）。该模型基于编码-解码机制，编码器学习到一个具有低秩结构的投影函数，用于将图像的视觉特征空间、语义描述空间以及标签进行连接；解码阶段重建原始视觉特征。并通过低秩嵌入，使得学习到的模型在预见未见类别时能共享已见类的语义信息，从而更好地进行分类。本研究在五个常见的数据集上进行实验，结果表明 LSSAE 的性能优于已有的零样本学习方法，是一种有效的零样本学习方法。

关键词： 图像分类；零样本学习；堆栈式自编码器；低秩嵌入
中图分类号： TP391 **文献标志码：** A **文章编号：** 1001-3695（2023）02-037-0539-05
doi：10.19734/j.issn.1001-3695.2022.06.0302

Zero-shot learning based on stacked semantic auto-encoder with low-rank embedding

Ran Ruisheng†，Dong Shuhong，Li Jin，Wang Ning
（*College of Computer & Information Science，Chongqing Normal University，Chongqing* 401331，*China*）

Abstract: In the field of image classification，existing methods such as deep learning require a large number of annotated samples for training and are unable to identify classes that do not appear in the training phase. Zero-shot learning tasks can effectively alleviate such problems. This study proposed a new zero-shot learning method，namely low-rank stacked semantic auto-encoder（LSSAE） based on stacked auto-encoder and low-rank embedding. The model was based on an encoding-decoding mechanism where the encoder learned a projection function with a low-rank structure for concatenating the visual feature space，the semantic space and the labels. It reconstructed the original visual features in the decoding stage. And the low-rank embedding enabled the learned model to share the semantic information of the seen classes when anticipating the unseen classes for better classification. Experiments were conducted on five common datasets in this study，and the results show that the proposed LSSAE outperforms existing zero-shot learning methods which is an effective zero-shot learning method.

Key words: image classification；zero-shot learning；stacked auto-encoder；low-rank embedding

0 引言

在图像识别与分类领域，诸如卷积神经网络（convolutional neural network，CNN）[1]、深度神经网络（deep neural network，DNN）[2]等模型都在大规模的图像识别和分类任务上获得了良好的表现。然而，这些模型都需要大量的训练样本，且只能对训练样本中出现过的类别进行分类。而在现实中，一方面收集大量的训练样本需要较高的成本，而且对于像长须鲸蓝鲸、中华鲟等处于濒危状态的动物，甚至都难以收集到样本；另一方面，由于传统图像分类方法无法识别在训练阶段未出现的类别，当有新的类别出现时往往需要重新训练一个模型。如果新的类别数据不断增多，每次都训练一个新模型所付出的代价会非常高。

为了解决上述问题，零样本学习（zero-shot learning，ZSL）被提出[3]，并受到广泛关注。ZSL 的任务是在训练阶段进行建模和学习，使模型能够识别在训练阶段未出现过的新类别。ZSL 的原理来自于人类识别新事物的机制。例如，如果一个小孩见过马、熊猫和老虎，他就知道马的外形，熊猫有黑白的毛，老虎有条纹等。这时，即使他没有见过斑马，但如果告诉他斑马的一些特征，如斑马有马的外形、有黑白相间的条纹等，当他见到斑马时，他就能准确地认出斑马这个新动物。图 1 给出了 ZSL 的任务示意图。

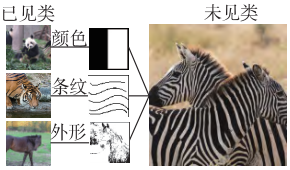

图 1 零样本学习任务
Fig. 1 Zero-shot learning tasks

具体来说，ZSL 是一个知识迁移的过程，旨在学习已见类与未见类之间的内在关系。已见类与未见类分别表示训练与测试的图像所属类别，两者的交集为空。ZSL 图像分类研究可以对训练数据集中没有出现的未知类进行预测，它在濒危动物的识别、故障检测、物体识别等诸多领域都具有很大的应用价值。

收稿日期：2022-06-29；修回日期：2022-08-18 基金项目：教育部人文社科规划项目（20YJAZH084）；重庆市技术创新与应用发展专项面上项目（cstc2020jscx-msxmX0190）；重庆市教委科学技术研究重点项目（KJZD-K202100505）

作者简介：冉瑞生（1976-），男（通信作者），教授，硕导，博士，主要研究方向为机器学习、计算机视觉（rshran@cqnu.edu.com）；董殊宏（1996-），男，硕士研究生，主要研究方向为机器学习、图像分类；李进（1998-），男，硕士研究生，主要研究方向为深度学习、计算机视觉；王宁（1994-），男，硕士，主要研究方向为计算机视觉.

现有的大多数 ZSL 方法主要通过以下几种方式实现训练与测试:

a) 通过属性预测[4]。直接属性预测(direct attribute prediction,DAP)与间接属性预测(indirect attribute prediction,IAP),直接或间接学习单个语义属性的分类器,通过属性预测实现已见类向未见类的知识迁移。

b) 通过学习视觉特征空间到语义空间的映射[5~12]。Akata 等人[5]提出属性标签嵌入的图像分类方法(attribute label embedding,ALE),该方法将图像从视觉特征空间映射到语义空间后学习一个兼容函数,用于衡量该图像与每个类别语义之间的匹配度,确保每张图像与所属的语义向量匹配度最高,测试只需要选中得分最高的类别标签即可。Frome 等人[7]提出一种深度视觉—语义嵌入模型(deep visual-semantic embedding model,DeViSE),首先预训练两个神经网络模型,一个用于语义向量或词向量,一个用于图像视觉特征,然后采用上述两个网络模型初始化提出来的模型。Zhang 等人[8]提出双重验证网络(dual-verification network,DVN)将图像特征投射到一个正交空间,使其与对应的语义属性有最大的相关性,并且与所有其他属性正交,最后计算出语义与标签之间的关系。Socher 等人[10]提出跨模态转移的零样本学习(cross-modal transfer,CMT),首先将图像通过神经网络模型映射到语义空间,接着将每个新的测试样本映射到这个语义空间中,确定是否在已见图像的流形上,如果图像是新的,即不在流形上,就在无监督的语义空间下进行分类。

c) 通过潜在空间进行学习[13~19]。直接将视觉特征空间映射到语义空间,由于两个不同域之间的语义差距较大,直接嵌入效果仍不够优秀。部分学者提出增加一个中间的共享空间,称为潜在空间。Xian 等人[14]提出潜在嵌入模型(latent embedding model,LATEM),通过学习多个映射函数,将视觉特征和语义信息映射到一个潜在子空间。其中不同的映射函数学习不同类别的视觉特征,对于不同的类别,模型选择兼容性最高的一组函数进行分类。Liu 等人[17]提出鉴别性双语义自编码器(discriminative dual semantic auto-encoder,DDSA),通过学习一个对齐空间构建两个双向映射自编码器,分类则将测试类别的语义投射到视觉空间,最后通过最近邻算法搜索分类。Wu 等人[18]提出视觉—语义联合优化模型(joint visual and semantic optimization,VSOP),将视觉特征空间与语义空间投射到一个潜在子空间,这样就可以提取这两个空间之间的关系进行分类。

然而,在 ZSL 中存在领域漂移的问题(domain shift)[20],即同一种语义属性信息在不同类别中也可能具有不同的视觉特征表达。另外,不同类别的视觉特征也可能传达相同或相近的语义信息,而上述方法都没有很好地关注到已见类与未见类中共享的某些语义信息。文献[21]提出的语义自编码器可以有效减少领域飘移的负面影响,其原因在于,虽然训练与测试图像的视觉特征分别来自可见类和未见类,但在解码阶段对视觉特征的重建约束对两者都有效,从而使得学习到的投影函数不容易受到领域漂移的影响。但是语义自编码器仅仅只关注到了图像的视觉特征到语义信息之间的映射关系,并没有很好地将图像的标签信息利用起来,不能建立图像特征到标签之间的映射关系。此外,已有研究表明低秩嵌入约束可以转移已见类与未见类的内在知识或共享特征[22],从而可以将已见类学习到的知识转移到未见类上。基于此,本文将低秩嵌入和语义自编码器相结合,提出基于低秩嵌入的堆栈语义自编码器的零样本学习方法(low-rank stacked semantic auto-encoder,LSSAE)。该方法主要贡献包含以下两方面:

a) 将堆栈式自编码器的思想用于 ZSL。该堆栈式自编码器的编码和解码过程分别含有两个隐藏层。编码器的第一层将视觉特征空间映射到语义空间,第二层将语义空间映射到标签空间。接下来,解码器通过标签信息与语义信息连续重建样本的原始特征信息。这样,在编码和解码过程,使模型学习到视觉特征空间到标签空间的映射关系。

b) 采用低秩嵌入的思想对模型进行约束,即模型尝试学习到一个有低秩结构的投影矩阵,以转移已见类别与未见类别的内在知识或共享特征。也就是说,通过这个方式可以为未见类别估算出更好的语义信息。

1 相关工作

本文的方法受到自编码器和低秩嵌入思想的启发,下面给出其简要说明。

1.1 自编码器(auto-encoder,AE)

自编码器是一种非监督学习算法,由编码器、隐藏层、解码器三部分组成。编码器将输入数据编码并映射到隐藏层,解码器则通过隐藏层解码重建原始输入。堆栈式自编码器(stacked auto-encoder)由编码器、多个隐藏层、解码器组成。输入数据经过多个隐藏层编码,逐层实现数据的抽象和特征提取。

语义自编码器(semantic auto-encoder)[21]是一种将样本语义和自编码器相结合的 ZSL 方法。该方法的编码器将图像从视觉特征空间投射到具有较低维度的语义空间,解码器则将语义空间投射回原始的视觉特征空间,旨在重建原始特征。测试任务则在两个空间中进行:

a) 通过编码器将新的测试样本投射到语义空间中,分类则只需要计算估计的语义与未见类的语义之间的距离。

b) 通过解码器将原始语义空间投射到特征空间,分类则计算测试样本的特征与估计的特征之间的距离。

1.2 低秩嵌入

由 ZSL 任务可以得知,训练与测试的数据样本分别取自已见类与未见类,两者虽然位于不同分布的特征空间,但它们却可能有相似的语义信息,当学习模型进行知识迁移时,就需要尽可能得到准确的未见类语义信息。由此 Ding 等人[22]提出将低秩嵌入(low rank embedding,LRE)的思想用于 ZSL 中。LRE 的思想主要假设未见类的语义信息与已见类的语义信息存在大部分共享,也就是不同类别间的语义信息能共享大部分属性值,这可在一个具有低秩结构的空间中确定。这样在对未见类的图像进行知识转移中就能更好地估计出潜在的语义信息。

2 提出的方法

2.1 符号和定义

假设有 c_s 个已见类别用于训练,其样本集合表示为 $D_s=\{X_s,A_s,Y_s\}$,已见类别标签为$\{1,\cdots,c_s\}$;有 c_u 个未见类别用于测试,其样本集合表示为 $D_u=\{X_u,A_u,Y_u\}$,未见类别标签为$\{1,\cdots,c_u\}$。$X_s\in\mathbb{R}^{d\times N_s}$和 $X_u\in\mathbb{R}^{d\times N_u}$分别表示训练图像和测试图像的特征向量集合,$d$ 是特征向量的维数,N_s 是训练样本数量,而 N_u 是测试样本数量。训练时的已见类别和测试时的未见类别的语义描述分别表示为 $A_s=\{a_i^s\}_{i=1}^{c_s}$,$A_u=\{a_j^u\}_{j=1}^{c_u}$。其中 $a_i\in\mathbb{R}^{k\times 1}$和 $a_j\in\mathbb{R}^{k\times 1}$是第 i 或 j 个类别的语义表示,k 是语义表示向量的维数。Y_s 和 Y_u 表示一组已见的和未见的类别标签,其中 $Y_s\cap Y_u=\varnothing$。$\boldsymbol{H}_s=\{h_1^s,h_2^s,\cdots,h_{N_s}^s\}\in\mathbb{R}^{k\times N_s}$是 N_s 个训练图像的语义描述矩阵,它们都来自于已见类别。

2.2 模型框架

如图 2 所示,在编码阶段,用编码器的第一层将输入的特征数据 X 投射到语义描述 H 中,即 $H=UX$;然后用类别语义描述矩阵 $\boldsymbol{A}$ 作为第二层编码器,来连接 H 和标签 Y,即 $Y=\boldsymbol{A}H$。在解码阶段,则通过$\hat{H}=\boldsymbol{A}^*\ Y$ 与$\hat{X}=U^*\ \hat{H}$连续重建原始特征。

为了便于计算，使用通过绑定权重 A 和 U 的转置即 A^{T} 和 U^{T} 作为解码器来重建原始视觉特征[23]，即 $\hat{H}=A^{\mathrm{T}}Y$，$\hat{X}=U^{\mathrm{T}}\hat{H}$。

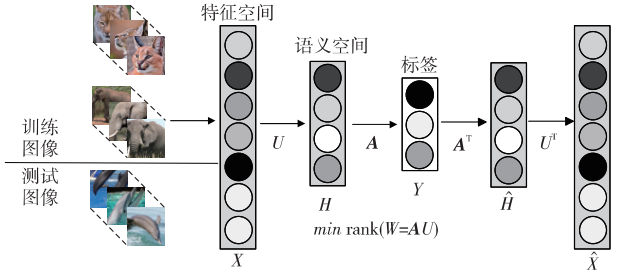

图 2　提出的模型

Fig. 2　The proposed model

本文的目标是最大限度地减少 $\hat{X}$ 和 X 之间的误差，这可以通过优化以下函数来实现：

$$\min_{U}\|X-U^{\mathrm{T}}A^{\mathrm{T}}AUX\|_F^2 \quad \text{s.t. } AUX=Y \tag{1}$$

其中：$\|*\|_F$ 表示矩阵的 Frobenius 范数，s.t. 表示为约束条件。

函数式(1)中，在约束条件 $AUX=Y$ 下难以求解。可将该约束条件放宽为 $\min\|AUX-Y\|_F^2$，这样函数式(1)可重写为

$$\min_{U}\|X-U^{\mathrm{T}}A^{\mathrm{T}}Y\|_F^2+\lambda\|AUX-Y\|_F^2 \tag{2}$$

函数式(2)的第一项可理解为解码器，第二项为编码器，λ 是用于平衡编码器和解码器之间的权重参数。

为了便于计算函数式(2)，令 $W=AU$，并将其表述为

$$\min_{W}\|X-W^{\mathrm{T}}Y\|_F^2+\lambda\|WX-Y\|_F^2 \tag{3}$$

文献[22]认为，未见类的语义向量与已见类的语义向量存在大部分共享标量，并可以在一个具有低秩结构的空间中确定。基于该观点，在堆栈式自编码器模型学习矩阵 W 的同时，对矩阵 W 加上低秩结构的约束，使得学习到的投影矩阵 W 在预测未见类别时能共享已见类的语义信息，这样在对未见类别进行知识转移时就能更好地估计出相应的语义信息，从而更好地进行分类。为此，将低秩约束条件 $\min\operatorname{rank}(W)$ 加到函数式(3)中，目标函数式(3)变为

$$\min_{W}\|X-W^{\mathrm{T}}Y\|_F^2+\lambda\|WX-Y\|_F^2+\beta\operatorname{rank}(W) \tag{4}$$

其中：β 是另一个平衡参数，$\operatorname{rank}(*)$ 表示矩阵的秩。式(4)中的秩最小化问题是一个著名的 NP 问题，有很多可行的解决方法。本文采用核范数 $\|W\|_*$ 来解矩阵的秩问题。在数学上，有以下等式：

$$\|W\|_*=\operatorname{tr}\left((W^{\mathrm{T}}W)^{\frac{1}{2}}\right)=\operatorname{tr}\left(W^{\mathrm{T}}(WW^{\mathrm{T}})^{-\frac{1}{2}}W\right) \tag{5}$$

由函数式(4)和(5)，在低秩约束下的新函数可以改写为

$$\min_{W}\|X-W^{\mathrm{T}}Y\|_F^2+\lambda\|WX-Y\|_F^2+\beta\operatorname{tr}(W^{\mathrm{T}}SW) \tag{6}$$

其中：$S=(\sqrt{WW^{\mathrm{T}}})^{-1}$。

2.3　模型求解

函数式(6)是一个标准的二次方程，其求解过程是一个凸优化的过程，且有全局最优解。为了求解它，只需要对函数取导，并令其导数为 0 即可。首先使用矩阵的迹重写函数式(6)。

a) 令 $f=\|X-W^{\mathrm{T}}Y\|_F^2$，通过矩阵的迹属性将其重写为 $f=\operatorname{tr}\left[(X-W^{\mathrm{T}}Y)(X-W^{\mathrm{T}}Y)^{\mathrm{T}}\right]$，由此得到导数表达式为

$$\frac{\mathrm{d}f}{\mathrm{d}W}=-2Y(X^{\mathrm{T}}-Y^{\mathrm{T}}W) \tag{7}$$

b) 令 $g=\lambda\|WX-Y\|_F^2$，同理可以将其重写为 $g=\lambda\operatorname{tr}\left[(WX-Y)(WX-Y)^{\mathrm{T}}\right]$，由此得到其导数表达式如下：

$$\frac{\mathrm{d}g}{\mathrm{d}W}=2\lambda(WX-Y)X^{\mathrm{T}} \tag{8}$$

c) 令 $z=\beta\operatorname{tr}(W^{\mathrm{T}}SW)$，其微分形式表达式可以表示为 $\mathrm{d}z=\beta\operatorname{tr}(\mathrm{d}W^{\mathrm{T}}SW+W^{\mathrm{T}}S\mathrm{d}W)$。对其求导则可以得到如下表达式：

$$\frac{\mathrm{d}z}{\mathrm{d}W}=2\beta(W^{\mathrm{T}}S)^{\mathrm{T}}=2\beta SW \tag{9}$$

然后将表达式(7)~(9)进行整合，可以得到下列等式：

$$(YY^{\mathrm{T}}+\beta S)W+W(\lambda XX^{\mathrm{T}})=(1+\lambda)YX^{\mathrm{T}} \tag{10}$$

接下来只需要令：$P=YY^{\mathrm{T}}+\beta S$，$Q=\lambda XX^{\mathrm{T}}$，$C=(1+\lambda)YX^{\mathrm{T}}$，则可以改写函数式(10)并得到如下的等式：

$$PW+WQ=C \tag{11}$$

形如式(11)的方程，是著名的 Sylvester 方程[24]。在基于 Python 的实验代码中，只需要引入 scipy 库包中的 solve_sylvester 方法即可求得有效解。接下来只需要计算 $U=(A_s+\eta E)^{-1}W$，这里的 η 表示一个极小正值，以防出现奇异性问题。

2.4　未见类预测

由于 A_u 在测试阶段是已知的，所以定义 $W_u=A_uU$，现在对于测试数据的优化问题就可以表述为

$$\min_{Y_u}\|X_u-W_uY_u\|_F^2+\lambda\|W_uX_u-Y_u\|_F^2+\beta\operatorname{rank}(W_u) \tag{12}$$

为了优化函数式(12)，同样计算它的导数并令其为 0，然后得到一个新的 Sylvester 方程为

$$P_2Y_u+Y_uQ_2=C_2 \tag{13}$$

其中：$P_2=WW^{\mathrm{T}}$，$Q_2=\lambda E$，$C_2=(1+\lambda)WX_u$。

因此在 Python 中可以很容易地得到 Y_u 的解，即

$$Y_u=\text{solve\_sylvester}(P_2,Q_2,C_2) \tag{14}$$

接下来，对 Y_u 的每一列进行归一化处理，并通过选择 Y_u 的每一行中最大值来预测测试实例，即

$$Y_u^*=\arg\max_{l}(Y_u^l),l\in[1,c_u] \tag{15}$$

3　实验分析

在 ZSL 领域，验证学习效果的常见数据集有五个，分别为 AWA1、CUB、SUN、aP&Y 和 AWA2。本章在五个数据集上进行了实验验证，给出了分类结果，另外以 AWA2 为例构建未见类混淆矩阵，可视化未见类语义分布情况，并对参数 λ 和 β 进行了分析，最后给出应用性分析。

3.1　数据集

Animal with Attributes(AWA1)[25]是一个中等规模的粗粒度数据集，共有 50 个动物类别，30 475 张图像，每个类别都由人类提供了 85 个语义描述。把其中的 40 个类别当作已见类，共 24 295 张图像用于训练，另外 10 个类别当作未见类，共 6 180 张图片用于测试。

Caltech-UCSD Birds-200-2011(CUB)[26]是一个中等规模的细粒度鸟类数据集，共有 200 个鸟类类别，11 788 张图像，每个类别都有 312 个语义描述。把其中的 150 个类别共 8 855 张图像用于训练，另外 50 个类别共 2 933 张图像用于测试。

SUN Attribute(SUN)[27,28]是一个中等规模的细粒度数据集，图像都与大型场景相关。其中包含了 717 种不同的场景类别，共 14 340 张图像，每个场景都标注了 102 个语义描述。把其中的 645 个场景共 12 900 张图像用于训练，另外 72 个场景共 1 440 张图像用于测试。

A Pascal and Yahoo(aP&Y)[29]是一个小规模粗粒度数据集，包含动物、目标、车辆共 32 个类别，每个类别都具有 64 个语义描述。其中 20 个类别用于训练，另外 12 个用于测试。

Animal with Attributes 2(AWA2)[30]使用了与 AWA 相同的 50 个动物类别，但包含了更多的图像，其数据划分方式与 AWA 一致。

表 1 中列出了各个数据集划分的统计信息。instance 表示图像样本数量；S-d 表示语义信息的维度；train 与 test 表示已见(训练)与未见(测试)类别的数量。

表 1 用于评估的数据集

Tab. 1 Benchmark datasets for evaluation

dataset	instances	S-d	train	test	dataset	instances	S-d	train	test
AWA1	30 475	85	40	10	aP&Y	15 339	64	20	12
CUB	11 788	312	150	50	AWA2	37 322	85	40	10
SUN	14 340	102	645	72					

3.2 评价标准

在 ZSL 中，采用的是每个测试类别的 top-1 精度的均值作为评价标准。其公式表达如下：

$$Acc = \frac{1}{\| c_u \|} \sum_{c=1}^{c_u} \frac{\text{样本正确预测数}}{\text{样本总数}} \tag{16}$$

其中：$\| c_u \|$ 为测试类别的总个数。这个衡量方式减少了对测试样本中数量较为稀少的类别的偏见，从而能更好地衡量模型性能。

3.3 实验对比

本文所有实验皆基于 Python 实现，所用的样本的视觉特征均为通过 GoogLeNet 提取的 1 024 维特征，语义信息均使用各个数据集提供的公开的人工标注的语义属性。

为了更好地验证 LSSAE 方法的有效性，将其与目前主流的相关方法进行了比较，分别为 direct attribute prediction (DAP) [4]，indirect attribute prediction (IAP) [4]，convex combination of semantic embeddings (ConSE) [3]，cross-modal transfer (CMT) [10]，semantic similarity embedding (SSE) [13]，latent embedding model (LATEM) [14]，attribute label embedding (ALE) [5]，deep visual-semantic embedding model (DeViSE) [6]，structured joint embedding (SJE) [7]，embarrassingly simple ZSL (ESZSL) [11]，synthesized classifiers (SYNC) [15]，generative framework for zero-shot learning (GFZSL) [16]，semantic auto-encoder (SAE) [21]，dual verification network (DVN) [8]，discriminative dual semantic suto-encoder (DDSA) [17]，visual and semantic optimization (VSOP) [18]。实验数据的比较结果如表 2 所示，以百分比列出对比方法与本文方法的准确率。其中“—”表示该方法在该数据集上并未报告结果。从表 2 可以看出，本文的 LSSAE 方法除了在 aP&Y 数据集上略低于 DDSA 之外，在其他四个数据集上都取得了最好的准确率。由此可知本文提出的 LSSAE 方法是一种有效的 ZSL 方法。

表 2 模型在数据集上的结果

Tab. 2 Experimental results on datasets

method	SUN	CUB	AWA1	AWA2	aP&Y	method	SUN	CUB	AWA1	AWA2	aP&Y
DAP	39.0	40.0	57.1	58.7	33.8	ESZSL	54.5	53.9	74.5	75.6	38.3
IAP	19.4	24.0	48.1	46.9	36.6	SYNC	56.3	55.6	72.7	71.2	23.9
ConSE	38.8	34.3	63.6	67.9	26.9	SAE	59.7	53.6	80.6	74.4	34.5
CMT	39.9	34.3	58.9	66.3	28.0	GFZSL	62.9	56.5	80.5	79.3	–
SSE	51.5	43.9	68.8	67.5	34.0	DVN	62.4	57.8	67.7	–	41.2
LATEM	55.3	49.3	74.8	68.5	35.2	DDSA	63.3	53.2	68.3	69.1	**46.1**
ALE	58.1	54.9	78.6	80.3	39.7	VSOP	60.8	52.6	75.2	74.1	45.3
DeViSE	56.5	52.0	72.9	68.6	39.8	LSSAE	**64.0**	**58.1**	**81.8**	**81.3**	45.2
SJE	53.7	53.9	76.7	69.5	32.9						

3.4 方法有效性分析

为了使分类结果更加直观，本文以 AWA2 数据集为例构建了其未见类的混淆矩阵，结果如图 3 所示。混淆矩阵的行与列分别代表真实的类别和预测的类别。从图中可以看出，所提方法可以有效地对未见类进行预测，特别是对“豹子”“座头鲸”几乎都可以准确地进行识别。由此可以说明，所提方法在对动物识别等应用场景中将会有很大的优势。

为了进一步验证 LSSAE 方法的有效性，通过训练完毕的模型，在将未见类测试样本从视觉空间投射到语义空间后，本文采用了 t-SNE [31] 算法来可视化 AWA2 数据集的未见类实例在语义空间中的分布情况，结果如图 4 所示。从中可以观察到，在图 4(a) 中，由 SAE 投射的语义信息各类比较分散，并未很好地聚集在一起，且其类别之间的差距较小。相比之下，在图 4(b) 中，本文方法学习的语义信息更容易单独地聚集在一起，并且几乎未发生重叠，类之间的差距也更加清晰。这意味着本文模型生成的语义信息的分布比 SAE 得到的更容易被分类，从而使得最终的分类结果有所提升。

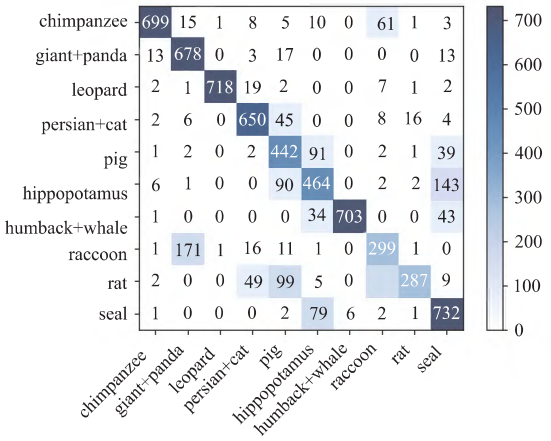

图 3 AWA2 中未见类的混淆矩阵

Fig. 3 Confusion matrix of the unseen classes on the AWA2 dataset

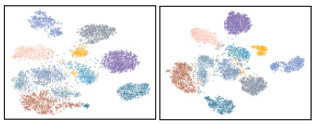

图 4 SAE 与 LSSAE 在 AWA2 上对未见类的 t-SNE 可视化

Fig. 4 T-SNE visualization of SAE and LSSAE of the unseen classes on the AWA2 dataset

3.5 参数分析

本文的 LSSAE 方法有两个参数，分别是 λ 和 β。通过控制变量法，对两个参数进行了分析。图 5(a) 和(b) 分别列出了参数 λ 和 β 在不同数据集上对模型的影响。对 λ 进行分析可以看出，在五个数据集上达到最高准确率时都是一个小范围值，当 λ 在值为 2 附近时表现最好。随着值增大，各个数据集准确率都有不同程度下降，其中 CUB 受影响程度最大，下降了 10% 左右，而 AWA 和 AWA2 受影响程度最小，在 2% 之内。当固定 λ 的值，对 β 进行分析，可以看出它对模型影响都较小，在五个数据集上的准确率都并未受到太大干扰，受影响的程度整体都不超过 1%。这说明在堆栈式自编码器模型结构下，λ 作为平衡编码器与解码器的平衡参数，模型对其相对更敏感，而对 β 不敏感，即性能指标基本稳定。总结分析，经验上可以将 λ 设置为 $1 \leqslant \lambda \leqslant 5$，而 $1 \leqslant \beta \leqslant 50$。

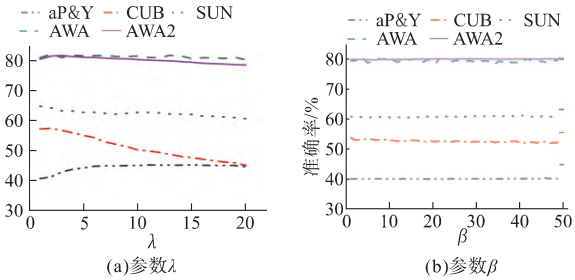

图 5 λ 与 β 参数分析

Fig. 5 Parametric analysis for λ and β

3.6 模型应用分析

前几节对所提方法进行了实验结果对比与有效性分析，证明了方法的有效性。下面以濒危动物识别的应用场景为例对所提方法的应用性进行分析说明，如图 6 所示。在训练阶段，本文使用大量常见的动物类别样本进行模型训练，求解得到投

影矩阵 U。注意，由于不容易获得中华鲟这种濒危动物的大量样本，在训练阶段没有中华鲟。接下来，如果有中华鲟的图像（假定事先并不认识中华鲟），对其进行特征提取，然后采用训练阶段得到的投影矩阵 U，即可进行预测分类。

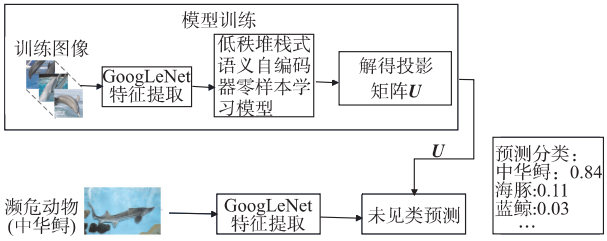

图 6　模型应用

Fig. 6　Application of the proposed model

4　结束语

基于堆栈式自编码器和低秩嵌入，本文提出了一种新的 ZSL 方法，名为基于低秩嵌入的堆栈语义自编码器（low-rank stacked semantic auto-encoder，LSSAE）。该方法的编码器将图像视觉特征映射到类别信息，而解码器则旨在精确地重建原始视觉特征信息。通过编码—解码机制，使得学习到的网络模型更能准确反映视觉特征与类别标签之间的映射关系。同时，通过低秩嵌入的约束，使得学习到的模型在预见未见类别时能共享已见类的语义信息，从而更好地进行分类。最后在 ZSL 领域的常见数据集上进行了实验，实验结果表明，LSSAE 优于其他先进的 ZSL 方法。另外，对 LSSAE 进行了有效性分析，验证了在动物识别场景下的优越性；还对两个超参数进行了分析，在不同的参数值下，对模型的影响基本都不大，这表明本文的模型具有很强的稳定性。但是，对于 ZSL 来说，目前总体准确率都不高，这是因为测试的类别都为未见类，底层视觉特征大多明显异于训练类别，训练阶段学习到的投影函数不能完全学习到类别视觉特征与语义信息之间的内在联系，证明在模型结构上还有更进一步的改进方案。其次，本文的模型要求人为标注不同类别之间的语义描述，存在主观性，未来的研究可以探索如何通过算法学习到类别语义信息之间更抽象的联系。最后，ZSL 存在的领域漂移等问题仍然存在，如何更有效地解决仍需要探索。

参考文献：

[1] Gu Jiuxiang，Wang Zhenhua，Kuen J，*et al*. Recent advances in convolutional neural networks [J]. Pattern Recognition，2018，77：354-377.

[2] Liu Weibo，Wang Zidong，Liu Xiaohui，*et al*. A survey of deep neural network architectures and their applications [J]. Neurocomputing，2016，234：11-26.

[3] Palatucci M，Pomerleau D，Hinton G E，*et al*. Zero-shot learning with semantic output codes [C] //Proc of the 22nd International Conference on Advances in Neural Information Processing Systems. 2009：1410-1418.

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

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

[6] Akata Z，Reed S，Walter D，*et al*. Evaluation of output embeddings for fine-grained image classification [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2015：2927-2936.

[7] Frome A，Corrado G S，Shlens J，*et al*. DeViSE：a deep visual-semantic embedding model [C] //Proc of the 26th International Conference on Neural Information Processing Systems. 2013：2121-2129.

[8] Zhang Haofeng，Long Yang，Yang Wankou，*et al*. Dual-verification network for zero-shot learning [J]. Information Sciences，2019，470：43-57.

[9] Norouzi M，Mikolov T，Bengio S，*et al*. Zero-shot learning by convex combination of semantic embeddings [EB/OL]. (2013). https://arxiv.org/abs/1312.5650.

[10] Socher R，Ganjoo M，Manning C D，*et al*. Zero-shot learning through cross-modal transfer [C] //Proc of the 26th International Conference on Neural Information Processing Systems. 2013：935-943.

[11] Romera P B，Torr P. An embarrassingly simple approach to zero-shot learning [C] //Proc of International Conference on Machine Learning. 2015：2152-2161.

[12] 张冀，曹艺，王亚茹，等. 融合 VAE 和 StackGAN 的零样本图像分类方法 [J]. 智能系统学报，2022，17(3)：593-601. (Zhang Ji，Cao Yi，Wang Yaru，*et al*. Zero-shot image classification method combining VAE and StackGAN [J]. CAAI Trans on Intelligent Systems，2022，17(3)：593-601.)

[13] Zhang Ziming，Saligrama V. Zero-shot learning via semantic similarity embedding [C] //Proc of IEEE International Conference on Computer Vision. Piscataway，NJ：IEEE Press，2015：4166-4174.

[14] Xian Yongqin，Akata Z，Sharma G，*et al*. Latent embeddings for zero-shot classification [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2016：69-77.

[15] Changpinyo S，Chao Weilun，Gong Boqing，*et al*. Synthesized classifiers for zero-shot learning [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2016：5327-5336.

[16] Verma V K，Rai P. A simple exponential family framework for zero-shot learning [C] //Proc of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin：Springer，2017：792-808.

[17] Liu Yang，Li Jin，Gao Xinbo. A simple discriminative dual semantic auto-encoder for zero-shot classification [C] //Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway，NJ：IEEE Press，2020：940-941.

[18] Wu Hanrui，Yan Yuguang，Chen Sentao，*et al*. Joint visual and semantic optimization for zero-shot learning [J]. Knowledge-Based Systems，2021，215：106773.

[19] 钟小容，胡晓，丁嘉昱. 基于潜层向量对齐的持续零样本学习算法 [J]. 模式识别与人工智能，2021，34(12)：1152-1159. (Zhong Xiaorong，Hu Xiao，Ding Jiayu. Continual zero-shot learning algorithm based on latent vectors alignment [J]. Pattern Recognition and Artificial Intelligence，2021，34(12)：1152-1159.)

[20] Fu Yanwei，Hospedales T M，Xiang Tao，*et al*. Transductive multi-view zero-shot learning [J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2015，37(11)：2332-2345.

[21] Kodirov E，Xiang Tao，Gong Shaogang. Semantic autoencoder for zero-shot learning [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2017：3174-3183.

[22] Ding Zhengming，Shao Ming，Fu Yun. Generative zero-shot learning via low-rank embedded semantic dictionary [J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2018，41(12)：2861-2874.

[23] Ranzato M A，Boureau Y L，LeCun Y. Sparse feature learning for deep belief networks [C] //Proc of the 20th International Conference on Neural Information Processing Systems. 2007：1185-1192.

[24] Bartels R H，Stewart G W. Solution of the matrix equation AX + XB = C [F4] [J]. Communications of the ACM，1972，15(9)：820-826.

[25] Lampert C H，Nickisch H，Harmeling S. Learning to detect unseen object classes by between-class attribute transfer [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2009：951-958.

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

[27] Patterson G，Hays J. Sun attribute database：Discovering，annotating，and recognizing scene attributes [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2012：2751-2758.

[28] Patterson G，Xu Chen，Su Hang，*et al*. The sun attribute database：beyond categories for deeper scene understanding [J]. International Journal of Computer Vision，2014，108(1)：59-81.

[29] Farhadi A，Endres I，Hoiem D，*et al*. Describing objects by their attributes [C] //Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE Press，2009：1778-1785.

[30] Xian Yongqing，Lampert C H，Schiele B，*et al*. Zero-shot learning：a comprehensive evaluation of the good，the bad and the ugly [J]. IEEE Trans on Pattern Analysis and Machine Intelligence，2018，41(9)：2251-2265.

[31] Van Der Maaten L，Hinton G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research，2008，9(86)：2579-2605.