doi:10.3969/j.issn.1672-1292.2025.04.004

基于变分自编码孪生网络的高分影像信息表征学习模型

孙　勇[1,2,3],胡方春[1],程千禧[1],顾程成[3],黄红鑫[2],谭文安[3]

(1.安徽大学资源与环境工程学院,安徽 合肥 230601)
(2.滁州学院实景地理环境安徽省重点实验室,安徽 滁州 239000)
(3.上海第二工业大学计算机与信息工程学院,上海 201209)

[**摘要**]　传统的监督分类方法在有效表征高分辨率遥感影像的语义信息方面存在困难,且通常需要大量准确的标记,但带有高质量标记的遥感数据相对匮乏. 针对遥感影像数据标记受限的问题,提出一种新的高分辨率影像表征学习模型,融合了变分自编码器与异步非对称结构的孪生神经网络,通过采用多个自监督增广的遥感影像作为当前样本的正例以实现平滑的表征. 在语义信息表征学习过程中,将孪生对比损失函数与变分自编码的标准证据下界结合进行优化,从而有效解决高分辨率遥感影像的无标记问题,并能捕获高分辨率影像的潜在语义信息. 所提算法在 SIRI-WHU、NWPU-RESISC45、UC Merced 及 AID 数据集上进行了实验,结果显示其性能优于最先进的自监督学习方法 SimSiam、MoCo、BYOL 和 SimCLR.

[**关键词**]　自监督学习,语义信息表征,高分辨率遥感影像,场景分类
[**中图分类号**]TP751;TP18　[**文献标志码**]A　[**文章编号**]1672-1292(2025)04-0028-09

Innovative Variational Autoencoding Twin Neural Network for Enhanced Representation Learning of High-Resolution Remote Sensing Images

Sun Yong[1,2,3],Hu Fangchun[1],Cheng Qianxi[1],Gu Chengcheng[3],Huang Hongxin[2],Tan Wenan[3]

(1.School of Resources and Environmental Engineering,Anhui University,Hefei 230601,China)
(2.Anhui Province Key Laboratory of Physical Geographic Environment,Chuzhou University,Chuzhou 239000,China)
(3.School of Computer and Information Engineering,Shanghai Polytechnic University,Shanghai 201209,China)

Abstract:Traditional supervised classification methods face significant challenges in effectively capturing semantic information from high-resolution remote sensing images,primarily due to the necessity for extensive and accurate labeling,coupled with the scarcity of high-quality labeled datasets. To address this issue,we propose a novel approach that integrates variational autoencoders with a twin neural network architecture. Our high-resolution image representation learning model utilizes multiple self-supervised augmented remote sensing images as positive examples for the current sample,thereby facilitating smoother representation learning. Furthermore,during the semantic representation learning process,we synergistically combine the twin contrastive loss function with the variational autoencoder framework. This dual optimization of the standard evidence lower bound enhances our ability to tackle the challenge of unlabeled remote sensing images while effectively capturing the latent semantic information inherent in high-resolution imagery. Experimental evaluations on benchmark datasets,including SIRI-WHU,NWPU-RESISC45,UC Merced,and AID,demonstrate that our proposed algorithm significantly outperforms state-of-the-art self-supervised learning methods,such as SimSiam,MoCo,BYOL,and SimCLR.

Key words:self-supervised learning,semantic information representation,high-resolution remote sensing images,scene classification

收稿日期:2024-09-27.
基金项目:安徽省教育厅重大重点科学研究项目(2022AH051113)、安徽省重点实验室开放基金资助项目(2022PGE003)、安徽省教育厅教学研究项目(2021jyxm1053).
通讯作者:孙勇,博士,副教授,研究方向:协同计算与地理空间人工智能. E-mail:ysun.nuaa@foxmail.com

高分辨率遥感影像数据的空间与纹理特征复杂多样,蕴含着丰富的语义信息. 由于大规模遥感影像数据具有高度复杂的几何特征和空间格局,如何有效表征高分辨率遥感影像的场景语义信息成为一个重要且具有挑战性的研究方向[1-2]. 为了解决这一问题,研究者们提出了多种方法,包括空间金字塔匹配核[3]、空间金字塔共现核[4]、KD-Tree[5]、稀疏编码[6],以及基于视觉词特征描述符学习表征影像语义信息[7]. 传统的特征提取方法依赖于专家知识的人工设计. 深度卷积神经网络(CNN)在处理高分辨率遥感影像方面展现出了优异的性能,通过逐级抽取影像样本数据中的特征,CNN 能够建立底层信息与高层语义之间的函数关系,从而自动提取语义特征. 数关系,卷积核自动提取语义特征.

遥感影像的语义信息表征是高分辨率遥感数据智能处理的关键要素. 基于深度学习的表征模型尤其依赖于大量高质量、结构化且有标签的数据. 然而,大部分高分辨率遥感影像缺乏高质量的标记,且人工标记的成本极高. 针对这一问题,自监督学习模型通过比较样本之间的相似性和差异性来学习特征表示,或者通过生成数据的方式来学习数据的分布,利用标签受限的大规模遥感影像数据学习语义表征向量,从而提高无标签遥感影像智能处理的准确性. 因此,基于自监督学习的深度表征方法已成为当前地理空间人工智能领域最具代表性的研究方向之一.

基于自监督学习的深度表征模型在遥感数字影像处理领域得到了广泛的应用,尤其是在语义分割[8]、目标检测[9]和影像分类[10]等任务中. 自监督学习模型主要分为以下几类:对比学习、生成模型、掩码预测、变换学习,以及任务导向的自监督学习模型[11]. 其中,对比学习通过比较样本之间的相似性和差异性来学习特征表示,常见方法包括 SimSiam[12]、SimCLR[13]、MoCo[14]、BYOL[15]以及视觉语言对比学习模型 CLIP[16],这些方法在遥感数据表征中得到了广泛应用[17-19]. 自监督生成模型则通过生成数据的方式来学习数据的分布,包括变分自编码器(VAE)[20]、生成对抗网络(GAN)[21]以及扩散模型[22]等,这类方法通常通过重构输入影像数据或生成新样本来进行训练. 掩蔽预测模型在输入数据中随机掩蔽部分信息,然后训练模型预测被掩蔽的部分[23-24]. SS-MAE[25]是一个典型的掩蔽自编码模型,通过掩蔽空间-光谱输入数据来进行自监督学习. 变换学习强调模型对数据变换的鲁棒性,通过对输入数据施加不同的变换(如旋转、缩放、裁剪等),训练模型学习不变性特征[11]. 任务导向的自监督学习模型则设计特定的任务来生成标签,例如影像的旋转预测、颜色化等,这些任务可以帮助模型学习到有用的特征表示[11].

在高分辨率遥感影像智能处理领域,基于自监督学习的深度表征方法取得了显著成效,但仍面临以下主要问题:(1)数据标注不足. 尽管自监督学习旨在减少对标签的依赖,但高分辨率遥感影像通常缺乏高质量的标注数据,这可能影响迁移模型的训练效果. (2)数据的多样性和复杂性. 遥感影像具有多样的地物类型和复杂的场景,如何设计有效的自监督任务以捕捉这些多样性和复杂性是一个挑战. (3)影像特征表示的有效性. 自监督学习的目标是学习有效的特征表示,但在遥感影像中如何确保学习到的特征能够有效区分不同的地物类型仍然是一个挑战.

针对上述问题,本文提出一种融合变分自编码器与孪生神经网络的高分辨率影像表征学习模型. 该模型结合多种自监督学习策略,以提高性能和泛化能力. 与现有的最先进方法不同,该模型利用 k 个正样本来学习遥感影像的语义表征向量. 具体而言,模型将 k 个自监督增强影像作为正样本进行平均值计算表征向量,从而创建平滑表示,这一方法能够有效减少噪声表示,显著提升特征提取的质量. 与此同时,在语义表征学习过程中,模型将孪生对比损失函数与变分自编码的标准证据下界相结合进行优化,这一策略有效提升了深度学习模型的整体性能. 在 SIRI-WHU、NWPU-RESISC45、UC Merced 和 AID 等不同高分辨率数据集上的实验结果表明,本文提出的深度表征学习模型能够快速有效地捕获高分辨率遥感影像中的潜在语义信息.

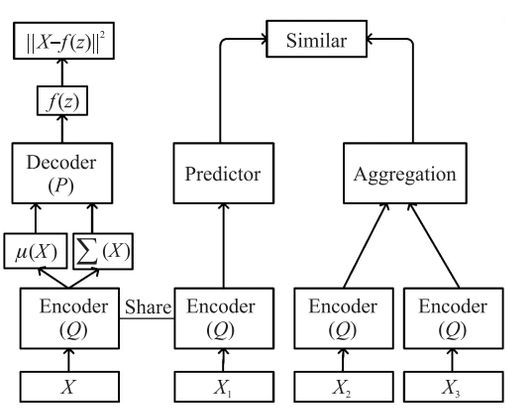

图 1　变分自编码孪生深度神经网络模型图

Fig. 1　Variational autoencoding twin deep neural network model

本文所提出的表征学习模型框架如图 1 所示.

1　原理与方法

1.1　变分自编码器模型

变分自编码深度学习模型主要由编码器(Encoder)和解码器(Decoder)两部分组成. 编码器和解码器均采用 ResNet 卷积神经网络进行设计与实现. 编码器将高分辨率遥感影像压缩到中间层,并采用低维的潜在向量空间进行表示. 解码器则根据低维特征向量重构原始的高分辨率遥感影像[20]. 变分自编码模型是一种基于自监督学习的深度神经网络模型,结合了自编码器深度学习思想与贝叶斯变分推断. 在训练过程中,模型生成一个与原始影像具有相同尺寸的影像输出,并将原始影像作为目标数据设计变分自编码损失函数 $\mathcal{L}=\|X-\mathrm{Dec}(\mathrm{Enc}(X;\psi);\theta)\|$. 这一过程不需要任何标记信息. 自编码器在重建原始遥感影像的过程中能够获取遥感数据的潜在语义特征,因此,高分辨率遥感影像的变分自编码可以作为下游智能处理任务的预训练方法.

1.1.1　遥感影像的编码器模型

遥感影像编码器是一个基于卷积神经网络的模型. 为了更高效地提取高分辨率遥感影像的潜在特征,将经典自编码器中的全连接网络层替换为卷积层,并对模型的部分结构和训练过程进行了相应的调整. 在编码过程 $Q(Z|X)$ 中,遥感影像样本被映射到潜在空间表征向量 $\boldsymbol{Z}$,具体定义如下:

$$\boldsymbol{Z}=\mathrm{Enc}(X;\psi)\backslash\sim Q(\boldsymbol{Z}|X;\psi). \tag{1}$$

式中,$\mathrm{Enc}(X;\psi)$ 是编码器函数,采用多层卷积神经网络实现,输入为高分辨率遥感样本点 $\boldsymbol{x}_i$,输出为潜在地理语义特征 z,ψ 表示卷积神经网络学习参数. 遥感影像编码器的目标是通过卷积神经网络学习编码函数,从而提取出低维特征向量,并有效挖掘潜在的地理语义信息. 为增强自编码器的性能,变分自编码器融合了统计推断理论,与将遥感影像映射到潜在空间中的固定编码不同,变分自编码器将遥感影像编码转化为统计分布的平均值和方差参数. 通常假设潜在特征向量服从正态分布,因此变分编码器函数可定义为:

$$(\boldsymbol{\mu},\log\boldsymbol{\sigma})=\mathrm{Enc}(X;\psi), \tag{2}$$

$$Q(\boldsymbol{Z}|X)=\mathcal{N}(z;\boldsymbol{\mu},\log\boldsymbol{\sigma}). \tag{3}$$

式中,$\boldsymbol{\sigma}$ 为统计分布的方差,$\boldsymbol{\mu}$ 为统计分布的平均值,统计分布的两个参数都是通过编码器神经网络学习获取的. 然后模型根据分布采用的潜在空间特征向量 z 进行学习.

1.1.2　潜在特征向量解码器模型

解码器的目标是将潜在语义特征重建为原始遥感影像,是编码器的逆过程,同样采取卷积神经网络学习解码器函数. 模型解码器的输入为潜在语义特征向量 z,输出则是与原始遥感影像数据尺寸大小相同的影像 $\boldsymbol{x}_i$,解码器的模型学习参数表示为 θ. 编码与解码过程的附加产物是下游高分遥感影像智能任务所需要的潜在语义特征向量. 模型的解码过程是通过 $P(X|\boldsymbol{Z};\theta)$ 将潜在语义特征 $\boldsymbol{Z}$ 映射到原始影像.

原始高分辨率影像样本数据 X 的概率分布为 $P(X)$,变分自编码器采用最大似然估计求解 $P(X;\theta)$,可将 $P(X;\theta)$ 等价转换为关于 X 的最大对数似然 $\log P(X;\theta)$ 的估计,求解模型可定义为:

$$\log P(X;\theta)=\mathbb{E}_{Z\sim Q(\boldsymbol{Z}|X;\psi)}\left[\log\frac{P(X,\boldsymbol{Z};\theta)}{P(X|\boldsymbol{Z};\theta)}\right]. \tag{4}$$

经过推导后,可得

$$\log P(X;\theta)=\mathop{\mathbb{E}}_{Z\sim Q(\boldsymbol{Z}|X;\psi)}\left[\log\frac{P(X,\boldsymbol{Z};\theta)}{Q(\boldsymbol{Z}|X;\psi)}\right]+\mathop{\mathbb{E}}_{Z\sim Q(\boldsymbol{Z}|X;\psi)}\left[\log\frac{Q(\boldsymbol{Z}|X;\psi)}{P(X|\boldsymbol{Z};\theta)}\right]. \tag{5}$$

式(5)的右边第二项是一个 $\mathcal{KL}$ 散度计算,$\mathcal{KL}$ 是恒大于或等于 0,故获取一个变分下界,即右边第一项变分自编码器的优化对象. 因此,变分自动编码损失可定义为:

$$\mathcal{L}_{\mathrm{VAE}}=-\mathbb{E}_{Z\sim Q(\boldsymbol{Z}|X;\psi)}[\log P(X|\boldsymbol{Z};\theta)]+\mathcal{KL}(Q(\boldsymbol{Z}|X;\psi)||P(\boldsymbol{Z})). \tag{6}$$

式中,$\mathcal{KL}(Q(\boldsymbol{Z}|X;\psi)||P(\boldsymbol{Z}))$ 表示 $\mathcal{KL}$ 散度,是一种量化概率分布 Q 与 P 之间差异. 在变分自编码器模型中,设计两个损失函数来学习影像语义表征参数,主要包括遥感影像重构损失 $\mathcal{L}_{\mathrm{rec}}$ 和正则化损失 $\mathcal{L}_{\mathrm{reg}}$. 遥感影像重构损失是为了迫使解码高分遥感影像匹配原始数据输入,采用重构 $\log P(\boldsymbol{x}_i|z;\theta)$ 似然函数评估潜在特征向量解码为原始遥感影像的学习效果. 变分自编码器的正化损失 $\mathcal{KL}(Q(\boldsymbol{Z}|X;\psi)||P(\boldsymbol{Z}))$,帮助学习具有

优良结构的潜在语义空间,同时可解决在遥感影像训练样本上的过拟合问题.

1.2　孪生深度神经网络

孪生深度神经网络是受自监督学习模型的启发,采用对比损失学习最大化不同的增广影像样本之间的相似性. 孪生深度神经网络由主干 ResNet[19] 和投影预测头 MLP(multilayer perceptron)组成,首先将影像 $\boldsymbol{x}$ 经过随机增广得到 $t_1(\boldsymbol{x})$ 与 $t_2(\boldsymbol{x})$,输入到编码器网络 Enc,通过编码器 Enc 提取特征,经过处理得到表征向量. 编码器在两个增广影像的网络之间共享权重,变换一个视图的输出并将其与另一个视图匹配,将两个输出向量表示为 $\mathrm{Enc}(t_1(\boldsymbol{x}))$ 与 $\mathrm{Enc}(t_2(\boldsymbol{x}))$.

孪生深度神经网络模型通过最大化相同样本的不同增广的语义特征之间的相似度,进而实现深度对比学习的目标,计算其中一个分支的投影向量和另一个分支的特征向量之间差异,作为损失函数,进行训练. 该模型采用非对称的结构,将投影头分别放在两个分支之后,进行损失函数计算. 在孪生网络分支采用交叉梯度更新优化的模式,在训练孪生网络分支 1 时,投影头放在分支 1 的编码器之后,计算损失函数,而分支 2 则停止回传梯度;在学习孪生网络分支 2 时,将分支 2 的投影头接入到分支 2 的编码器之后,并计算损失函数,而分支 1 则停止回传梯度. 整个学习过程都是采用交叉梯度更新的方式进行.

孪生神经网络损失函数计算公式为:

$$\mathcal{L}_{\mathrm{Twin}}=\mathbb{E}_{(\boldsymbol{x},t_1,t_2)}\left[\left\|\mathrm{renorm}(P_\gamma(\mathrm{enc}(t_1(\boldsymbol{x}))))-\mathrm{sg}(\mathrm{renorm}(\mathrm{Enc}(t_2(\boldsymbol{x}))))\right\|_2^2\right], \tag{7}$$

式中,$\boldsymbol{x}$ 从高分辨率影像样本数据 X 取样,$\mathrm{remorn}(\cdot)$ 表示 L_2 正则化函数,可定义为:

$$\mathrm{renorm}(\boldsymbol{v})=\frac{\boldsymbol{v}}{\max(|\boldsymbol{v}|_2+\zeta)}, \tag{8}$$

式中,ζ 为调节参数. 孪生神经网络的目标是最大化正样本之间的表征向量相似性. 同一批量样本中与当前样本特征不同的被认为是负样本,通常还需要最小化负样本表征向量与当前样本之间的相似性.

当前的孪生深度神经网络,例如 SimSiam 和 SimCLR,是基于单个增广高分影像视图进行训练目标损失函数,导致训练不够稳定. 通过对高分影像进行多个增广视图训练,可使得训练更加稳定,进而实现更快地表征学习. 因此,本文的孪生深度神经网络基于多正样本聚合表征影像,其公式如下:

$$\mathrm{feat}_{\mathrm{agg}}=\frac{1}{K}\sum_{k=1}^{K}\mathrm{Enc}(t_k(x)), \tag{9}$$

式中,$\mathrm{feat}_{\mathrm{agg}}$是多个正样本聚合提取的特征表示,$K$ 表示多个正样本.

本文将变分自编码器与基于多正样本聚合表征的孪生神经网络进行融合,设计了一种多任务同时训练的集成深度学习框架. 在该框架中,孪生神经网络的对比学习机制与变分编码器的损失度量函数相结合. 在语义信息表征学习过程中,采用自监督学习技术对当前样本进行增广,生成多个遥感影像作为当前样本的正例,从而实现平滑表征. 此外,将对比损失函数与变分自编码的标准证据下界同时优化,有效提升了深度学习模型的性能. 融合变分自编码器与孪生深度学习框架的损失函数可定义为:

$$\mathcal{L}_{\mathrm{Total}}=\mathcal{L}_{\mathrm{Twin}}+\alpha\times\mathcal{L}_{\mathrm{VAE}}, \tag{10}$$

式中,α 是多任务目标函数的缩放因子;$\mathcal{L}_{\mathrm{Twin}}$是孪生神经网络的对比损失函数;$\mathcal{L}_{\mathrm{VAE}}$是变分自编码损失函数. 变分自编码孪生深度神经网络采用随机梯度下降(SGD)优化器进行预训练,并使用余弦权重衰退机制来设置学习率. 在编码网络中,投影多层感知器(MLP)部分的每个全连接层后接批归一化(BN)层,而输出的全连接层(FC 层)后不使用 ReLU 激活函数,隐含层的 FC 维度设置为 2 048,MLP 包含 3 个全连接层. 本模型采用 ResNet 作为主干网络,进行无监督预训练,随后采用监督学习方式冻结主干网络以训练分类器.

2　高分辨率遥感影像语义信息表征实验

本实验从高分辨数据集中提取表征向量用作学习分类器的输入,进行多层感知器、逻辑回归、微调测试以及最近邻搜索等实验. 表征实验收集了 4 种常用的遥感影像数据集,其中 SIRI-WHU、NWPU-RESISC45 和 UC Merced 数据集拥有 256×256 像素的高分辨率遥感影像,AID 数据集拥有 600×600 像素的高分辨率遥感影像.

UC Merced 土地利用数据集包含海滩、丛林、港口、跑道、停车场、网球场等 21 类,共计 2 100 张土地利

用遥感影像,其中每种场景包含 100 张 256×256 像素的影像,每张影像在 RGB 色彩空间的像素分辨率达到 0.3 m 左右. UC Merced 数据通常用于航空影像分类,如图 2 所示.

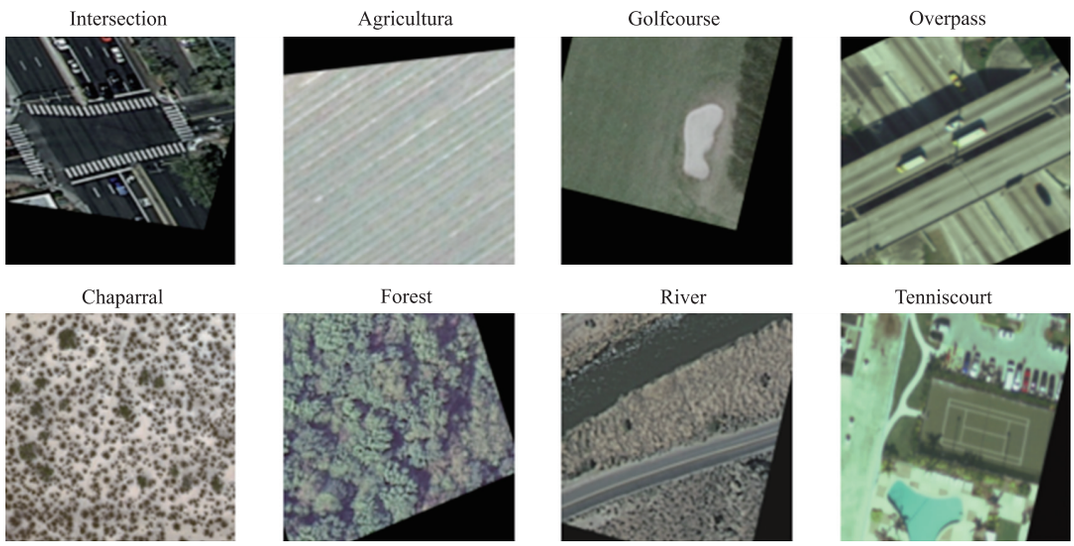

图 2　UC Merced 遥感影像示例

Fig. 2　UC Merced scene image samples

SIRI-WHU 数据集来源自谷歌地球,主要包括中国城市区域的闲置用地、农场、商业区、港口、工业区、草地、立交桥、停车场、池塘、居民区、河流以及水体等,共计 12 种不同类别,包含 200 张 256×256 像素的影像,空间分辨率为 2 m,如图 3 所示.

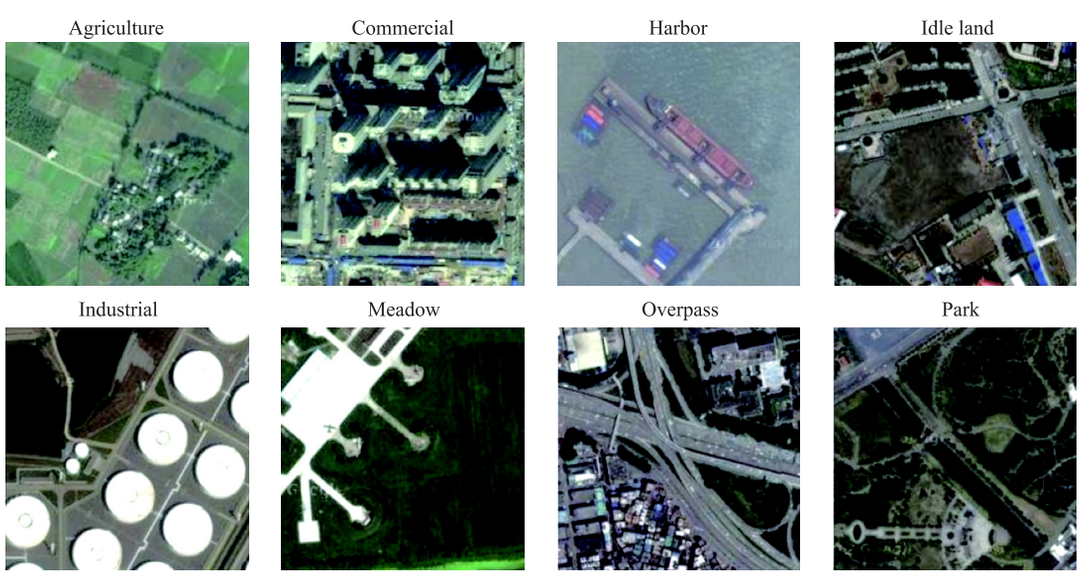

图 3　SIRI-WHU 场景影像示例

Fig. 3　SIRI-WHU scene image samples

NWPU-RESISC45 数据集是用于遥感影像场景分类规模最大的公开数据集合,所含信息丰富,包含了棒球场、篮球场、飞机、机场、海滩、桥梁、丛林、教堂、圆形农田、云、商业区等 45 类特定的场景,每类场景有 700 张影像,共计 31 500 张,涵盖了全球 100 多个国家和地区,属于较大规模的影像数据集. NWPU-RESISC45 场景影像在空间分辨率、视点、照明、背景和遮挡方面存在着差异,具有类内差异性和类间相似性的特点. 其影像像素为 256×256,空间分辨率在 0.3 到 30 m 范围内,如图 4 所示.

AID 数据集是从谷歌地球收集整理得到的大规模高分辨率遥感影像数据集,包含教堂、商业、密集住宅、沙漠、农田、森林、工业、草地、中型住宅、山地、公园、停车场、机场、裸地、棒球场、海滩、桥梁、中心游乐场、池塘、港口、火车站、度假村、河流、学校、稀疏住宅、广场、体育场、高架桥等,共计 30 种遥感场景类

型. 每一类场景包含 200 到 400 幅样本影像,共 10 000 幅样本影像. AID 具有更大的类内距离、更小的类间距离、更大的规模等特征,随着空间分辨率的提高,场景的几何结构越来越清晰,给表征与分类任务带来了更大的挑战,如图 5 所示.

图 4　NWPU-RESISC45 场景影像示例

Fig. 4　NWPU-RESISC45 scene image samples

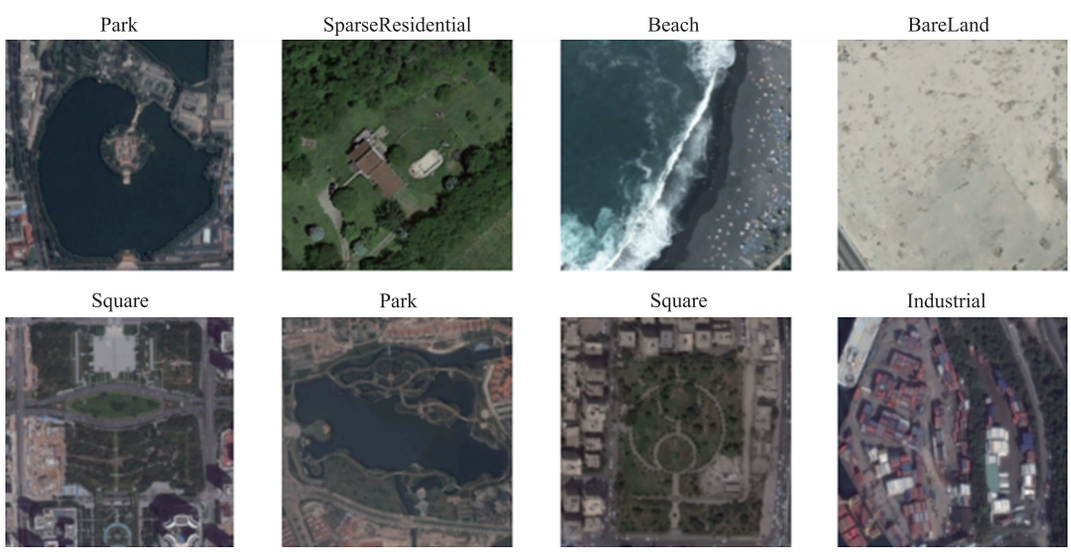

图 5　AID 场景影像示例

Fig. 5　AID scene image samples

2.1　实验参数设置

本文实验采用 Python 作为编程语言,基于 PyTorch 和 Scikit-learn 框架进行,集成开发环境为 Visual Studio Code. 实验所用的 CPU 为 Intel(R)Xeon(R)CPU@2.20GHz,GPU 为 Tesla P100-PCIE-16GB. 在训练过程中,将数据集随机划分为训练集和测试集,比例为 8∶2. 实验中采用 ResNet-18 作为骨干网络架构. 各项实验参数定义如下:Epoch 表示迭代次数,设置为 100;LR 表示学习率,设置为 0.06;OPT 表示优化器,选用 SGD;MO 表示动量,设置为 0.9;BatchSize 表示批量大小,设置为 128;InputSize 表示影像输入分辨率,设置为 256.

2.2　实验结果分析

2.2.1　基于高分影像信息表征的机器学习分类实验

本节实验采用多层感知器(MLP)和逻辑回归(LR)等机器学习方法,以验证高分辨率影像语义表征模型的显著有效性. 首先,从 UC Merced 高分辨率遥感影像数据集中提取特征,并将这些特征作为输入应用

于 3 个机器学习分类器,以评估目标任务测试集的分类率. 实验随机选取 UC Merced 遥感影像数据集中的训练数据,训练样本与测试数据的比例为 8∶2. 实验评估过程重复进行 10 次,并对分类率进行平均. 实验结果如表 1 所示.

表 1　MLP 与 LR 机器学习分类测试实验

Table 1　Classification experiments on UC Merced with MLP and LR

模型名称	MLP-Top1/%	MLP-Top5/%	LR-Top1/%	LR-Top5/%	Average/%
MoCo	78.253 93	95.873 01	79.206 34	97.936 50	87.817 45
SimCLR	79.365 08	95.714 28	78.095 23	96.031 74	87.301 58
SimSiam	79.523 80	97.460 31	81.746 02	97.936 50	89.166 65
BYOL	78.095 23	96.825 39	79.682 54	97.301 58	87.976 18
Proposed	82.380 95	98.888 86	83.333 33	99.047 61	90.912 68

注:Proposed 表示本文提出变分自编码孪生网络的高分影像语义表征模型.

本文提出的高分辨率遥感影像语义表征模型在实验中显示出最佳性能. 在表征向量的分类任务中,使用多层感知器(MLP)和逻辑回归(LR)分类器的平均分类率达到了 90.91%,在 5 种自监督分类器中表现最佳,优于 MoCo、SimCLR、SimSiam 和 BYOL 等自监督学习模型的结果,且该算法不依赖于任何人工标注成本,并在遥感数据集上有效应用,平均分类率显示出最佳效果. 特别值得注意的是,其他自监督学习算法的最高准确率为 89.16%,而本文所提模型达到了 90.91%.

同时,实验选择了 UC Merced 遥感影像数据集,比较所提出的自监督表征学习模型与最新的视觉语言对比学习模型 CLIP[18] 的分类性能. 其中,当视觉语言模型 CLIP 的图像编码器选择 ResNet50 时,其 Top-1 分类准确率为 73.14%;当选择 ViT-B/16 时,Top-1 分类准确率为 83.62%. 相比之下,本文提出的自监督表征学习模型采用了相对轻量的 ResNet18 作为图像编码器,其 Top-1 分类准确率达到了 83.33%,与 CLIP 在使用 ViT-B/16 时的性能相当. 这一结果表明,合理设计的自监督学习策略能够有效地学习到有价值的遥感图像视觉表征,成功弥补了模型复杂度较低的不足,并展现出良好的分类性能.

2.2.2　遥感影像语义表征模型微调实验

本文的微调测试采用了 UC Merced、SIRI-WHU、NWPU-RESISC45 和 AID 4 个遥感影像数据集. 在所有高分辨率遥感数据集中,80%的数据用于训练,其余 20%用于测试. 与分类测试类似,实验中使用每种方法预训练的权重参数作为目标任务的初始权重,唯一的例外是最后一个线性分类层. 实验结果如表 2 所示.

表 2　语义表征模型的微调测试实验

Table 2　Fine-tune experiments of semantic representation model

遥感影像数据集	Fine-tune Top 1/%	Fine-tune Top 5/%	遥感影像数据集	Fine-tune Top 1/%	Fine-tune Top 5/%
UC Merced	85.873 01	99.206 34	NWPU-RESISC45	86.825 39	98.571 42
SIRI-WHU	88.472 22	99.583 33	AID	82.733 33	97.299 99

本实验采用微调测试方法,以验证多个经典高分辨率影像的语义表征模型的显著有效性. 从表 2 可知,在标记数据有限的情况下,本文所提出的基于遥感影像语义表征模型的迁移学习方法在 UC Merced、SIRI-WHU、NWPU-RESISC45 和 AID 数据集上均表现出良好的效果.

2.2.3　高分影像信息表征的最近邻查询实验

在高分辨率影像的最近邻查询中,本文采用训练好的自编码孪生网络的编码器作为特征提取器,以测量所提出算法学习的表征质量. 实验在 UC Merced 数据集的语义表征向量空间内进行,针对给定样本数据 q,搜索距离最近的影像:

$$\mathrm{Neighbor}(q)=\arg\min_{x\in\mathcal{X}}\mathrm{Dist}(q,\boldsymbol{x}), \tag{11}$$

式中,$\mathrm{Dist}(q,\boldsymbol{x})$ 表示查询样本 q 与遥感影像表征向量 $\boldsymbol{x}$ 之间的距离. 实验随机选择了 3 个查询样本数据,并搜索最近的 2 个影像数据. 该实验结果如图 6 所示.

图 6 中,Query Image 表示给定的查询高分辨率影像样本,d 表示最近邻影像特征与 Query Image 样本特征之间的距离,距离越小越好. 高分影像信息表征的最近邻搜索实验结果表明,本文所提模型具有良好的表征质量,能够有效地搜索出相似的遥感影像,适用于标签受限情况下的遥感影像搜索系统.

综上所述,所提出的遥感影像语义表征模型在 UC Merced、SIRI-WHU 和 AID 数据集上进行了分类、微

调和最近邻搜索实验,取得了令人满意的效果. 因此,在标记数据受限的情况下,该基于变分自编码孪生网络的表征学习算法是有效的.

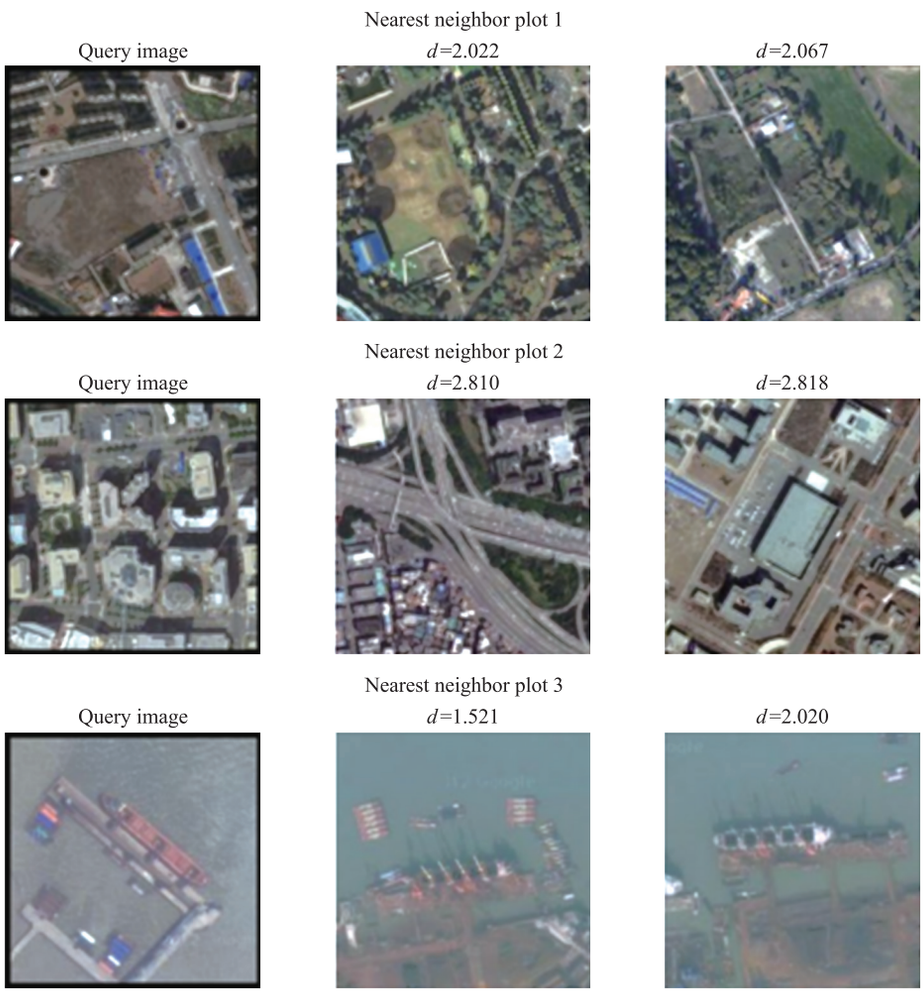

图 6　UC Merced 场景影像最近邻搜索示例

Fig. 6　Examples of UC Merced scene images nearest neighbors searching

3　结论

本文提出了一种面向高分辨率遥感影像语义信息表征的变分自编码孪生深度神经网络学习模型. 该模型借鉴了深度自监督学习的表征建模方法,通过使用多个增广的遥感影像来增加训练数据的多样性,从而确保学习过程的稳定性. 在语义信息表征学习过程中,本文将对比损失函数与变分自编码的标准证据下界相结合进行优化,这一策略有效提升了深度学习模型的性能,为高分辨率遥感信息的智能处理提供了有力支持. 在 UC Merced、SIRI-WHU、NWPU-RESISC45 和 AID 数据集上进行的实验表明,所提出的模型在性能上优于当前最先进的自监督学习方法. 具体而言,模型在表征向量上使用多层感知器(MLP)与逻辑回归(LR)分类器,平均分类准确率达到 90.91%,在 5 种分类器中表现最佳,高于 MoCo、SimCLR、SimSiam 和 BYOL 等自监督学习模型所获得的最佳结果. 此外,所提出的遥感影像语义表征模型在 UC Merced、SIRI-WHU、NWPU-RESISC45 和 AID 数据集上进行微调和最近邻搜索实验时,也取得了良好的效果. 在遥感影像标记受限的情况下,本文所提出的变分自编码孪生网络学习的表征算法展现出良好的有效性,为高分辨率遥感影像的智能处理提供了新的思路和方法.

[参考文献](References)

[1]　何小飞,邹峥嵘,陶超,等. 联合显著性和多层卷积神经网络的高分影像场景分类[J]. 测绘学报,2016,45(9):1073-

1080.
[2] 李冠东,张春菊,王铭恺,等. 卷积神经网络迁移的高分影像场景分类学习[J]. 测绘科学,2019,44(4):116-123.
[3] VAILAYA A,FIGUEIREDO M A T,JAIN A K,et al. Image classification for content-based indexing[J]. IEEE Transactions on Image Processing,2001,10(1):117-130.
[4] LAZEBNIK S,SCHMID C,PONCE J. Beyond bags of features:Spatial pyramid matching for recognizing natural scene categories[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). New York,USA:IEEE,2006.
[5] GUEGUEN L. Classifying compound structures in satellite images:A compressed representation for fast queries[J]. IEEE Transactions on Geoscience and Remote Sensing,2015,53(4):1803-1818.
[6] CHERIYADAT A M. Unsupervised feature learning for aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2014,52(1):439-451.
[7] SCHMIDHUBER J. Deep learning in neural networks:An overview[J]. Neural Networks,2015,61:85-117.
[8] MUHTAR D,ZHANG X L,XIAO P F. Index your position:A novel self-supervised learning method for remote sensing images semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing,2022,60:4411511.
[9] MA S T,HOU B,WU Z T,et al. Automatic aug-aware contrastive proposal encoding for few-shot object detection of remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing,2023,61:5615211.
[10] FENG Z X,SONG L L,YANG S Y,et al. Cross-modal contrastive learning for remote sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2023,61:5517713.
[11] CHEN X L,HE K M. Exploring simple siamese representation learning[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville,USA:IEEE,2021.
[12] CHEN T,KORNBLITH S,NOROUZI M,et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the 37th International Conference on Machine Learning. Online:JMLR,2020.
[13] HE K M,FAN H Q,WU Y X,et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,USA:IEEE,2020.
[14] GRILL J B,STRUB F,ALTCHÉ F,et al. Bootstrap your own latent:A new approach to self-supervised learning[C]//Proceedings of the 34th International Conference on Neural Information Processing System. Vancouver,Canada:NIPS,2020.
[15] JEAN N,WANG S,SAMAR A,et al. Tile2Vec:Unsupervised representation learning for spatially distributed data[J]. Proceedings of the AAAI Conference on Artificial Intelligence,2019,33(1):3967-3974.
[16] LUO Y,LEONG C T,JIAO S H,et al. Geo-Tile2Vec:A multi-modal and multi-stage embedding framework for urban analytics[J]. ACM Transactions on Spatial Algorithms and Systems,2023,9(2):10.
[17] WANG Y,ALBRECHT C M,BRAHAM N A A,et al. Self-supervised learning in remote sensing:A review[J]. IEEE Geoscience and Remote Sensing Magazine,2022,10(4):213-247.
[18] ZHANG J Y,HUANG J X,JIN S,et al. Vision-language models for vision tasks:A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2024,46(8):5625-5644.
[19] LIU F,CHEN D L,GUAN Z Q Y,et al. RemoteCLIP:A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing,2024,62:5622216.
[20] GIRIN L,LEGLAIVE S,BIE X,et al. Dynamical variational autoencoders:A comprehensive review[J]. Foundations and Trends in Machine Learning,2021,15(1/2):1-175.
[21] SONG B Z,LIU P,LI J,et al. MLFF-GAN:A multilevel feature fusion with GAN for spatiotemporal remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing,2022,60:4410816.
[22] YANG L,ZHANG Z L,SONG Y,et al. Diffusion models:A comprehensive survey of methods and applications[J]. ACM Computing Surveys,2023,56(4):105.
[23] HE K M,CHEN X L,XIE S N,et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans,USA:IEEE,2022.
[24] LIN J Y,GAO F,SHI X C,et al. SS-MAE:Spatial-spectral masked autoencoder for multisource remote sensing image classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2023,61:5531614.
[25] WANG Y,HERNÁNDEZ H H,ALBRECHT C M,et al. Feature guided masked autoencoder for self-supervised learning in remote sensing[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2024,18:321-336.

[责任编辑:严海琳]