引用格式：NI Kang, ZHAO Yuqing, CHEN Zhi. Multi-scale Convolutional Neural Network Driven by Sparse Second-order Attention Mechanism for Remote Sensing Scene Classification[J]. Acta Photonica Sinica, 2022, 51(6):0610004
倪康,赵雨晴,陈志.稀疏二阶注意力机制驱动的多尺度卷积遥感图像场景分类网络[J].光子学报,2022,51(6):0610004

稀疏二阶注意力机制驱动的多尺度卷积遥感图像场景分类网络

倪康[1,2],赵雨晴[3],陈志[1]
(1 南京邮电大学 计算机学院,南京 210023)
(2 江苏省大数据安全与智能处理重点实验室,南京 210023)
(3 首都经济贸易大学 管理工程学院,北京 100070)

摘 要:地面目标尺度信息不同及场景图像复杂的空间分布和纹理信息导致基于CNNs的场景分类算法分类效果欠佳。针对以上问题,从深度特征学习角度出发,提出一种稀疏二阶注意力机制驱动的多尺度卷积神经网络。首先,在主干网之后引入多尺度卷积层以获取地面目标不同尺度信息目标的特征表述,将组卷积嵌入多尺度卷积层以降低计算复杂度;其次,在分析基于一阶和二阶统计量的注意力机制优势之后,提出一种稀疏二阶注意力机制,以增强不同尺度卷积特征的通道信息可判别性。该注意力机制的稀疏性可在确保场景分类性能的同时,有效降低二阶统计量的特征维度;最后,将多尺度卷积层与稀疏二阶注意力机制嵌入端到端网络训练。在AID和NWPU45数据集上的实验表明:本文所提网络可提升场景分类准确率;同时,通过热力图结果对比和消融实验,验证了稀疏二阶注意力机制和所提各网络层的有效性。

关键词:遥感图像;卷积神经网络;特征可判别性;多尺度卷积;注意力机制
中图分类号:TP751 **文献标识码**:A **doi**:10.3788/gzxb20225106.0610004

0 引言

遥感图像解译是遥感图像信息处理的关键内容之一[1-4]。随着遥感技术的不断发展,高分辨率遥感图像的数量和成像质量均得到了快速增长和提升,这使得传统基于人工目视的遥感图像解译工作不能够满足正常需求[5]。近年来,深度学习等相关理论知识的发展和应用,使得基于大规模数据量的遥感图像解译工作在解译速度和精度上均有明显提升。因此,遥感图像解译得到了国内外专家和学者的广泛关注。

遥感图像场景分类通过分析单幅高分辨率场景图像中的内容并赋予其相应的类别标签,是遥感图像解译的重要内容之一,现已广泛应用于交通管制、灾情预测等领域[6]。但由于场景图像地面目标的多样性和空间信息的复杂性使得对场景图像的场景内容理解极具挑战性。

近年来,深度学习理论的快速发展为高分辨率遥感图像场景分类提供了有效途径。相比于传统的基于手工设计的特征描述子,深度特征表述在特征鲁棒性和泛化性上的优势[7],使得基于深度学习的高分辨率遥感图像场景分类迅速成为遥感图像信息处理领域的研究热点之一。卷积神经网络(Convolutional Neural Networks, CNNs)在遥感图像场景分类领域取得了较优的分类效果[8]。目前,按照特征学习的方式,基于CNNs的遥感图像场景分类方法可以分为:基于预训练CNNs特征提取和基于端到端CNNs特征学习的场景分类算法。

基金项目:国家自然科学基金(No. 62101280),江苏省自然科学基金(No. BK20210588),南京邮电大学引进人才科研启动基金(No. NY220135)
第一作者:倪康(1991—),男,讲师,博士,主要研究方向为遥感图像处理、SAR图像处理。Email: tznikang@163.com
通讯作者:陈志(1978—),男,教授,博士,主要研究方向为软件工程、无线传感网、物联网、数据挖掘。Email: chenz@njupt.edu.cn
收稿日期:2022-02-10;**录用日期**:2022-03-22

基于预训练CNNs特征提取的遥感图像场景分类方法是一种将现有的ImageNet等自然图像数据集上训练的深度神经网络作为特征提取器，提取遥感图像场景图像的深度特征，继而训练分类器，以完成场景分类。CHENG G等[9]利用预训练的AlexNet、GoogleNet和VGGNet-16网络作为特征提取器，并将所提取到的特征向量作为视觉词袋模型(Bag of Visual Words, BoVW)的输入，以此提升其特征的可辨别性。为了利用多层网络的深度特征，HE N J等[10]提出了一种多层堆叠的协方差池化网络(Multilayer Stacked Covariance Pooling, MSCP)，该网络提取预训练的CNNs网络中的多层深度特征向量，并采用协方差池化的方法进行获取其二阶特征统计信息，以此完成高分辨率遥感图像场景分类。为了充分顾及深度卷积特征中级特征表述和特征冗余对场景分类效果的影响，NI K等[11]提出一种基于中级深度特征学习的遥感图像场景分类算法。该算法利用一种可学习的多层激励局部约束仿射子空间编码-卷积神经网络框架(Learnable Multilayer Energized Locality Constrained Affine Subspace Coding-Convolutional Neural Network, MELASC-CNN)进行深度特征学习。YANG Z等[12]提出一种多尺度特征融合遥感图像场景分类算法，该算法通过输入不同尺度的遥感图像，提取预训练CNNs中的卷积层与全连接层特征，继而进行特征降维操作，将降维后的特征向量输入多核支持向量机(Multi-Kernel Support Vector Machine, MKSVM)完成场景分类。上述算法仅仅将CNNs作为特征提取器，利用特征降维、特征融合等算法提升深度特征的可判别性，以提高高分辨率遥感图像场景分类的准确率。但该类算法忽略了CNNs的特征学习能力，故而限制了其在遥感图像场景分类精度上的提升空间及其泛化能力。

基于端到端CNNs特征学习的场景分类算法突破了上述瓶颈，并取得了较好的场景分类效果。LU X Q等[13]在考虑深度特征聚合策略之后，提出一种卷积特征聚合编码网络，以此获取遥感场景图像的类别标签。该网络未顾及深度语义特征对遥感图像特征描述的影响，因此，LI R Y等[14]在充分利用多级和多尺度深度特征的同时，将深度语义特征信息融入特征金字塔网络，以自动学习场景图像判别特征表述。上述网络可针对不同的深度特征向量进行端到端特征聚合，并可提升模型的泛化能力。为了增强深度特征的可辨别性，注意力机制的引入成为研究热点。

深度学习中的注意力机制可以有效且自动地进行特征选择，应用较为广泛的注意力模块有：SENet(Squeeze-and-Excitation Networks)[15]、CBAM(Convolutional Block Attention Module)[16]、GCNet[17]和ECANet(Efficient Channel Attention)[18]等。上述注意力机制模块在图像分类、目标检测等领域取得了较好的效果，但该类模块大多是情况下是利用深度特征的一阶特征统计量进行相关性学习，这种方式在一定程度上限制了其表述能力。因此，基于二阶统计量的注意力模块相继出现。GAO Z L等[19]将全局二阶池化模块(Global Second-order Pooling, GSoP)嵌入卷积神经网络并得到不错的效果。BRYAN X等[20]提出一种非局部二阶注意力网络(Second-order Non-local Attention Network, SONA-Net)，该网络通过二阶特征统计量获取特征的长距离依赖。虽然上述基于二阶特征统计量的注意力模块在相关领域已取得较好的效果，但通过二阶统计量所得到的深度特征相关性的特征维度较高。例如：卷积特征向量$\boldsymbol{X}=\mathrm{R}^{H\times W\times C}$，其二阶统计量特征维度为$C^2$。因此，在不损害深度特征统计性能的情况下，约减特征维度至关重要。

综上所述，本文针对遥感场景图像地面目标尺度信息不同及场景图像复杂的空间分布和纹理信息导致基于CNNs的场景分类算法分类效果欠佳的问题，从深度特征学习角度切入，提出一种稀疏二阶注意力机制驱动的多尺度卷积神经网络(Multi-scale Convolutional Neural Network Driven by Sparse Second-Order Attention Mechanism, MCNN-SSAM)。本文在主干网之后引入金字塔卷积以提取场景图像的多尺度深度特征，减弱遥感场景图像地面目标尺度信息不同对场景图像特征信息描述的影响；另外，引入稀疏二阶注意力模块对多尺度卷积中不同尺度卷积层的通道信息进行通道选择，以此提高深度特征向量的可判别性。

1 MCNN-SSAM原理

1.1 MCNN-SSAM网络架构

本文提出的MCNN-SSAM包含以下几个部分：主干网、金字塔卷积模块、稀疏二阶注意力模块和Softmax分类层，如图1所示。另外，为了更好地验证主干网卷积层特征提取效果，图2给出了不同网络层(VGG-M)所学习到的特征图可视化结果。

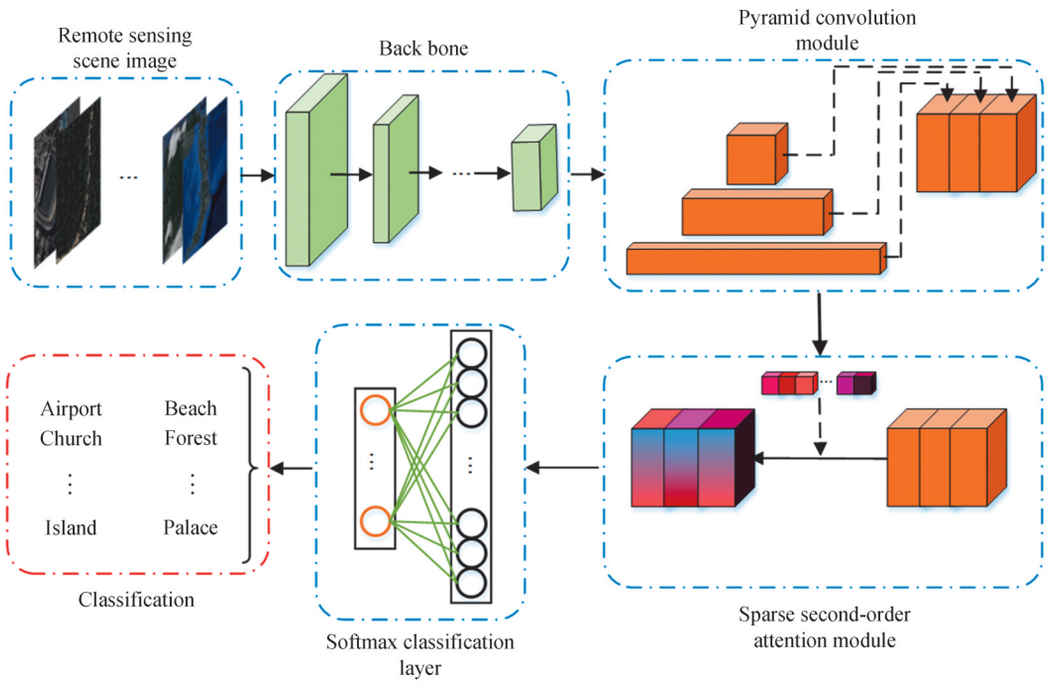

图1　MCNN-SSAM网络结构图

Fig.1　The architecture of MCNN-SSAM

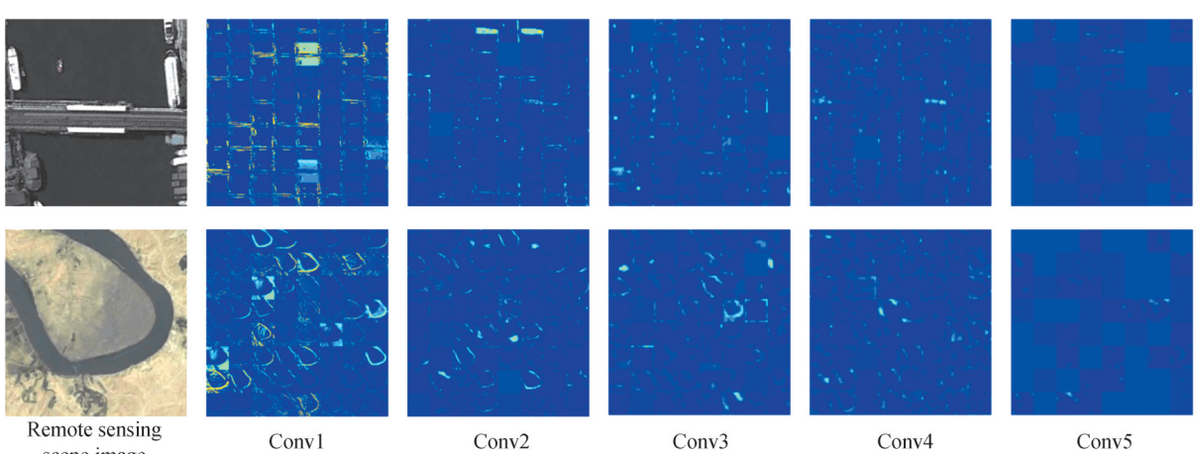

图2　不同网络层所学习到的特征图可视化结果

Fig.2　Visualization results of feature maps learned by different network layers

从图1可以看出：浅层的卷积层（如Conv1）所得到特征图的特征激活区域较广，部分通道的特征图刻画不精确，例如：在桥梁场景图像中，部分特征图关注在水域等区域，忽略了桥梁区域；另外，浅层卷积层特征图大多表现为边缘、轮廓等特征信息；随着网络层数的加深，特征图激活区域更加稀疏，特征也表现的更加抽象，且具有更高层的语义信息，例如在河流场景图像中，河流边缘、轮廓等特征信息逐渐稀疏，Conv5所提取的特征图激活区域少且集中，特征表现形式更加抽象，更具可辨别性。因此，本文截取VGG-16最后一个卷积层（Conv5_3）之前所有的网络层作为主干网。另外，金字塔卷积中包含三个不同尺度的卷积操作，稀疏二阶注意力模块可自动学习不同尺度卷积层的通道信息并进行通道选择，最后嵌入Softmax分类层完成网络的端到端训练。下文将详细阐述所提网络的各个部分。

1.2　金字塔卷积模块

金字塔卷积网络结构块采用3个级别的卷积并联而成。每个卷积都有四个参数(h, w, c, g)，分别代表卷积核的高、宽、通道数和组卷积中组的数目[21]。这里，卷积核的高和宽设置为$[3, 5, 7]$，即$h_1 = w_1 = 3$，

$h_2=w_2=5$，$h_3=w_3=7$，每一级卷积核的通道数均相等（按照经典CNNs通道数目的设置，c通常设置为2的指数幂，本文中$c=512$）。为了保证不同卷积所得到的特征图大小相同，该结构块中步长（stride）均设置为1，填充（padding）设置为$h//2$，// 为整数除法，返回不大于结果的一个最大的整数。这样通过卷积输出特征图的计算公式，即可得到同等大小的特征图输出。

另外，经典卷积操作的参数量和计算量均来自于卷积核计算。假设输入特征图通道数为c_{in}，输出特征图通道数为c_{out}，输出特征图大小为$[H, W]$，卷积核大小为k，则参数量大小P与浮点数计算量F为

$$\begin{aligned} P &= k^2 \cdot c_{\text{in}} \cdot c_{\text{out}} \\ F &= k^2 \cdot c_{\text{in}} \cdot c_{\text{out}} \cdot (H \cdot W) \end{aligned} \tag{1}$$

在输出特征图保持一致的情况下，因卷积核大小的不同，与单尺度卷积层相比，金字塔卷积网络结构块的参数量和计算量明显提高。为了降低金字塔卷积网络结构块的参数量和计算量，考虑到经典卷积操作中，卷积操作针对每个通道的特征图都进行类似于全连接计算方式的特征图卷积，这种计算方式直接影响了卷积核的参数量与计算量。金字塔卷积网络结构块按照通道数目进行分组，再进行卷积操作，这样每个分组内的特征图进行独立的卷积操作，模块的参数量和计算量都会随着c_{in}和c_{out}的降低而明显降低。若分组数目$g=1$，即演变为经典的卷积操作，但分组数目过多也会影响到特征学习的效果[21]。金字塔卷积网络结构块中g_1、g_2和g_3拟选定为相等的参数，另外根据CNNs通道数的设置，将其设置为2的指数幂，本文设置为4。最后，通过BN（Batch Normalization）层和ReLU（Rectified Linear Unit）非线性激活层对输出特征图进行非线性建模，增加模型的表达能力和特征泛化能力，降低网络过拟合现象的发生。

1.3 稀疏二阶注意力模块

本文所提出的稀疏二阶注意力模块结构图如图3所示。对于金字塔卷积模块的输出特征图$G\in\mathrm{R}^{H\times W\times C}$，利用$1\times1$卷积层进行通道降维，可得其降维的特征图为$H\times W\times C'$。继而利用FBC模块（Factorized bilinear coding）进行稀疏二阶统计量的计算。

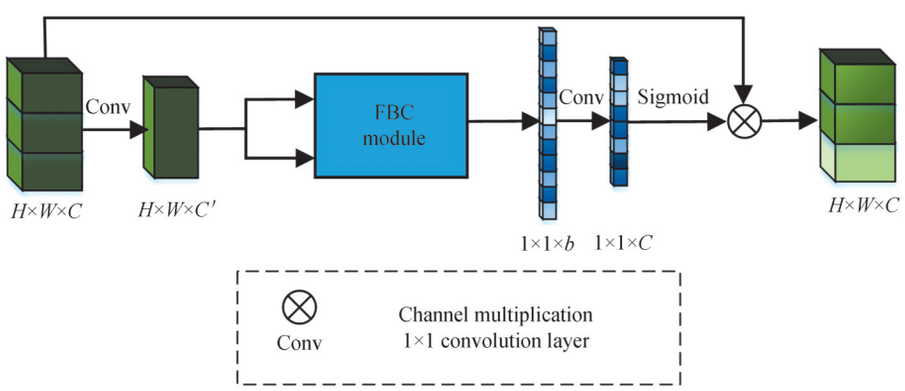

图3 稀疏二阶注意力模块结构图

Fig.3 The architecture of sparse second-order attention module

FBC模块的核心思想是从深度特征向量特征对$(\boldsymbol{m}_i, \boldsymbol{n}_j)$中学习到一个具有$b$个原子的字典$\boldsymbol{A}$，其中每一个原子均可以分解为低秩矩阵$\boldsymbol{X}_l\boldsymbol{Y}_l^{\mathrm{T}}$[22]。此时，编码系数$\boldsymbol{c}_s$可通过式（2）计算。

$$\min_{\boldsymbol{c}_s}\left\| \boldsymbol{m}_i\boldsymbol{n}_j^{\mathrm{T}} - \sum_{l=1}^{k}\boldsymbol{c}_s^l\boldsymbol{X}_l\boldsymbol{Y}_l^{\mathrm{T}} \right\|^2 + \omega\left\|\boldsymbol{c}_s\right\|_1 \tag{2}$$

式中，ω为可调参数，$s=1,2,\ldots,C$，$\|\cdot\|_1$为l_1范数算子，$\boldsymbol{c}_s^l$表示编码系数的第l个元素。$\boldsymbol{X}_l\in\mathrm{R}^{p\times v}$，$\boldsymbol{Y}_l^{\mathrm{T}}\in\mathrm{R}^{v\times q}$，$q$是超参数，且$v\ll p$。从式（2）可以看出，FBC模块可以通过LASSO（Least Absolute Shrinkage and Selection Operator）算法求解，即

$$\boldsymbol{c}_s = \mathrm{sign}(\boldsymbol{c}_s')\odot\max\left(\mathrm{abs}(\boldsymbol{c}_s') - \omega/2, 0\right) \tag{3}$$

式中，$\boldsymbol{c}_s'=\boldsymbol{Q}\left(\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{m}_i\odot\boldsymbol{Y}'^{\mathrm{T}}\boldsymbol{n}_j\right)$，$\odot$为Hadamard积，且$\boldsymbol{Q}\in\mathrm{R}^{b\times vb}$为固定的二值矩阵，$\boldsymbol{X}'$和$\boldsymbol{Y}'$是通过低秩矩阵$\boldsymbol{X}$和$\boldsymbol{Y}$计算得到的，其目的是为了降低运算复杂度。这里，$\boldsymbol{X}'$和$\boldsymbol{Y}'$表示为

$$\begin{cases} \boldsymbol{X}_l = \boldsymbol{I}\left(\left(\boldsymbol{p}_l \boldsymbol{1}_{vb}^{\mathrm{T}}\right) \odot \boldsymbol{X}^{\mathrm{T}}\right) / v \\ \boldsymbol{Y}_l = \left(\boldsymbol{I}\boldsymbol{Y}^{\mathrm{T}}\right) / v \end{cases} \tag{4}$$

式中，$\boldsymbol{1}_{vb}^{\mathrm{T}}$和$\boldsymbol{I}$为全1向量和矩阵[23]，$\boldsymbol{p}_l$为$\boldsymbol{P}$的第$l$列，其中$\boldsymbol{P}$定义为

$$\boldsymbol{P} = \left(\left(\boldsymbol{Q}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{Q}^{\mathrm{T}} \odot \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Q}^{\mathrm{T}}\right)\right)^{-1}\boldsymbol{Q}\right)^{\mathrm{T}} \tag{5}$$

通过上述求解，利用FBC模块求得的稀疏二阶注意力特征向量可以表示为

$$\boldsymbol{F} = \max\left\{\boldsymbol{c}_s\right\}_{s=1}^{b} \tag{6}$$

式中，$\boldsymbol{F} \in \mathrm{R}^{1\times 1\times b}$是通过最大化操作，遍历字典$\boldsymbol{A}$中每个原子聚合得到的，且$b << C^2$。此时，稀疏二阶注意力模块的计算表达式为

$$\boldsymbol{G}' = \sigma\left(\mathrm{Conv}^{1\times 1}(\boldsymbol{F})\right) \otimes \boldsymbol{G} \tag{7}$$

式中，$\mathrm{Conv}^{1\times 1}(\bullet)$为$1\times 1$卷积操作，其目的是通过降维操作，完成对深度卷积特征$\boldsymbol{G}$通道信息的自动学习，建模金字塔卷积特征空间域通道维度之间的相互依赖性，有效地增强有价值特征信息的特征响应，抑制无价值特征信息的特征响应。$\sigma(\bullet)$为sigmoid激活函数，$\otimes$为特征图通道乘法。稀疏二阶注意力模块在从二阶特征统计量获取多尺度深度特征通道之间相关性的同时，顾及二阶特征统计量的特征冗余性，可以得到更好的特征增强效果。最后，引入逐像素归一化层增强特征的可判别性，主要包括符号平方根归一化层和l_2归一化层，具体计算公式为

$$\begin{cases} \boldsymbol{G}_i^{j'} = \mathrm{sign}\left(\boldsymbol{G}_i^{j'}\right)\sqrt{\left|\boldsymbol{G}_i^{j'}\right| + \kappa} \\ \boldsymbol{G}_i^{j'} = \sqrt{\sum_{j=1}^{N}\left(\boldsymbol{G}_i^{j'}\right)^2 + \kappa} \end{cases} \tag{8}$$

式中，$\boldsymbol{G}_i^{j'}$为$\boldsymbol{G}'$特征向量的第i行第j列个特征描述子，$\mathrm{sign}(\bullet)$为符号函数，即：当$\boldsymbol{G}_i^{j'} > 0$时，$\mathrm{sign}\left(\boldsymbol{G}_i^{j'}\right) = 1$；当$\boldsymbol{G}_i^{j'} = 0$时，$\mathrm{sign}\left(\boldsymbol{G}_i^{j'}\right) = 0$；当$\boldsymbol{G}_i^{j'} < 0$时，$\mathrm{sign}\left(\boldsymbol{G}_i^{j'}\right) = -1$。$\kappa$为一个小整数，以保证算式有意义。

2 实验结果与分析

2.1 实验数据集

为了验证所提网络的有效性，利用两个应用较为广泛且数据量规模较大的遥感图像场景分类数据集进

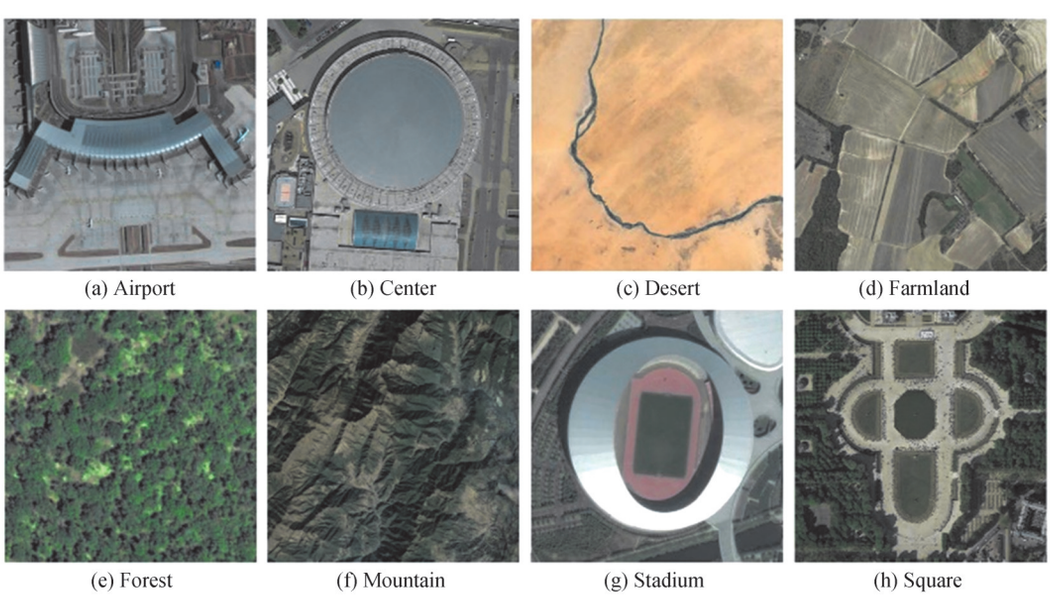

图4 AID场景实例

Fig.4 Sample charts of AID dataset

行验证实验[24-25]。AID(Aerial Image Dataset)数据集中的场景图像来自于谷歌地球,其中包括10 000幅图像尺寸为600像素 × 600像素的场景图像,空间分辨率为1~8 m,共分为30个场景类别(如图4所示):机场、海滩、桥梁、山脉、森林等。

NWPU45(NWPU-RESISC45 dataset)数据集是一个由沙漠、篮球场、湖泊、岛屿、火车站等45个类别的场景图像所构成的,共包括31 500幅图像尺寸为256像素 × 256像素的场景图像,空间分辨率为0.2~30 m。NWPU45数据集场景示例如图5所示。

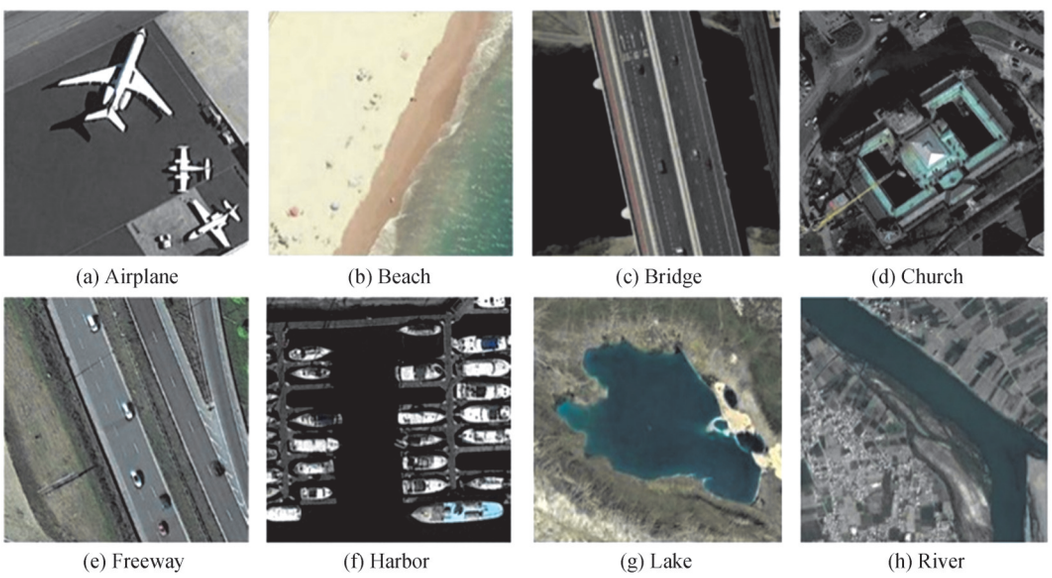

图5 NWPU45场景实例

Fig.5 Sample charts of NWPU45 dataset

2.2 实验设置

选定VGG-16网络作为MCNN-SSAM的主干网,利用Adam优化器进行网络端到端训练,相关参数设置如下:初始学习率0.01、权重衰减系数0.001、数据批次大小(batch size)32、动量0.9,其他参数设置将在2.3节中讨论。本文利用PyTorch深度学习框架进行网络搭建与实验,硬件平台为:GPU: NVIDIA GeForce GTX 8G 1070 Ti和RAM: 32 GB。

2.3 参数分析

本文所提MCNN-SSAM中稀疏二阶注意力模块包的两个重要参数:字典中的原子数量b和低秩矩阵参数v,对所提网络的场景分类性能有较大影响。故图6给出了不同的b和v下,MCNN-SSAM的AID遥感图像场景数据集上的场景分类总体精确度(Overall Accuracy, OA)。这里随机选取数据集中20%场景图像作为训练集,10%作为验证集,70%作为测试集。

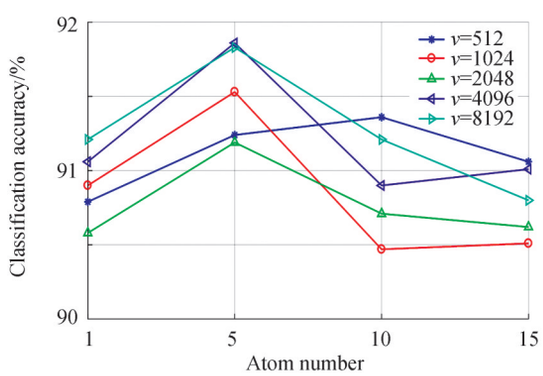

图6 b和v对MCNN-SSAM网络性能的影响

Fig.6 Influence of b and v on MCNN-SSAM

从图6可以看出，不同的低秩矩阵参数和原子数量对网络性能会有不同程度的影响。当原子数量不断增加时，大多数情况下，网络分类准确率是先升后降，这种现象的产生多数情况下与所构建网络的复杂度和训练样本数量有关。特别地，并不是低秩矩阵参数和字典中的原子数量越大，网络性能越优。过大的低秩矩阵参数和原子数量会导致稀疏二阶注意力模块计算复杂度的增加，且过多的原子数量亦不能较好地建模金字塔卷积特征空间域通道维度之间的相互依赖性。通过上图可以看出，当$v=8\,192$和$v=4\,096$在$b=5$时，可以得到91.83%和91.86%的场景分类准确率。在综合考虑算法计算复杂度的同时，MCNN-SSAM在后续实验过程中将低秩矩阵参数和原子数量设置为4 096和5。

2.4 与其他相关算法对比

为了对比本文所提MCNN-SSAM与其他相关网络的有效性，本小节利用AID和NWPU45公开数据集进行验证实验。在AID数据集中，随机选取20%和50%场景图像作为训练集，其余作为测试集；在NWPU45数据集中，随机选取10%和20%场景图像作为训练集，剩余的90%和80%作为测试集。为保证实验结果的可靠性，所有实验均进行5次。所对比的相关算法包括AlexNet、VGG-16、SAFF、MSCP和CapsNet遥感图像场景分类算法。表1给出了相关实验结果，实验结果中，“+”前面的数字为5次实验结果场景分类准确率的均值，“+”后面的数字为5次实验结果标准差大小，该数值可以衡量模型的稳定性。

表1 不同算法效果对比

Table 1 Performance comparison of other related algorithms

	AID		NWPU45	
	20%	50%	10%	20%
AlexNet	86.86±0.47	89.53±0.31	76.69±0.21	79.85±0.13
VGG-16	86.59±0.29	89.64±0.36	76.47±0.18	79.79±0.15
SAFF	90.25±0.29	93.83±0.28	84.23±0.19	87.86±0.14
MSCP	91.52±0.21	94.42±0.17	85.33±0.21	88.93±0.14
CapsNet	91.63±0.19	94.74±0.17	85.08±0.13	89.18±0.14
MCNN-SSAM	**91.86±0.09**	**94.98±0.11**	**86.67±0.10**	**90.61±0.11**

AlexNet和VGG-16网络是经典的CNNs，在遥感图像场景分类方面，该类网络可得到一定的场景分类效果。SAFF[26]和MSCP这两种算法均采用了多尺度特征聚合的方式，不同的是：SAFF中所运用的是基于自注意力机制的特征融合方式；MSCP采用二阶深度特征统计量描述遥感图像场景特征。从实验结果可以看出，在不同场景分类数据集和训练集比例的情况下，与AlexNet和VGG-16网络相比，SAFF和MSCP均有不同程度的提升（约为3.5%～9.0%）。CapsNet[27]将胶囊网络（CapsNet）与CNNs有效结合，并以此提升遥感图像场景分类性能，并得到了与MSCP性能相当的场景分类准确率。本文所提MCNN-SSAM不仅利用金字塔卷积进行多尺度深度特征抽取，而且引入稀疏二阶注意力模块获取深度多尺度特征通道之间相关性，在顾及二阶特征统计量特征冗余性的同时，可以得到更好的特征增强效果。因此，MCNN-SSAM可获得较好的场景分类效果，与基准网络VGG-16相比，有5.0%～11.0%的性能提升；与性能较好的CapsNet相比，也有0.20%～1.50%的性能提升。

为了更好地展示每一类遥感图像场景图像的分类效果，图7和图8分别给出本文所提MCNN-SSAM在AID数据集20%训练集比例和NWPU45数据集20%训练集比例下的场景分类混淆矩阵。混淆矩阵中坐标系数字为数据集中每个类别名称，均按照其英文首字母顺序排序，横坐标为预测类别，纵坐标为真实类别。

从该混淆矩阵可以看出，大部分场景类别可得到80%以上的遥感图像场景分类准确率。值得注意的是：部分场景类别，如AID数据集中的山脉、高架桥，NWPU45数据集中的丛林、海冰等场景均达到了98%～99%的场景分类准确率。另外，AID数据集中娱乐场、NWPU45数据集中宫殿的场景分类准确率在70%以下，造成该现象的原因可能是由于娱乐场和AID数据集中的公园场景类间相似性较大，因此，造成23%的娱乐场类别误分为公园。同理，NWPU45数据集中宫殿与教堂有较大的类间相似性，故有15%的宫殿类别误分为教堂。

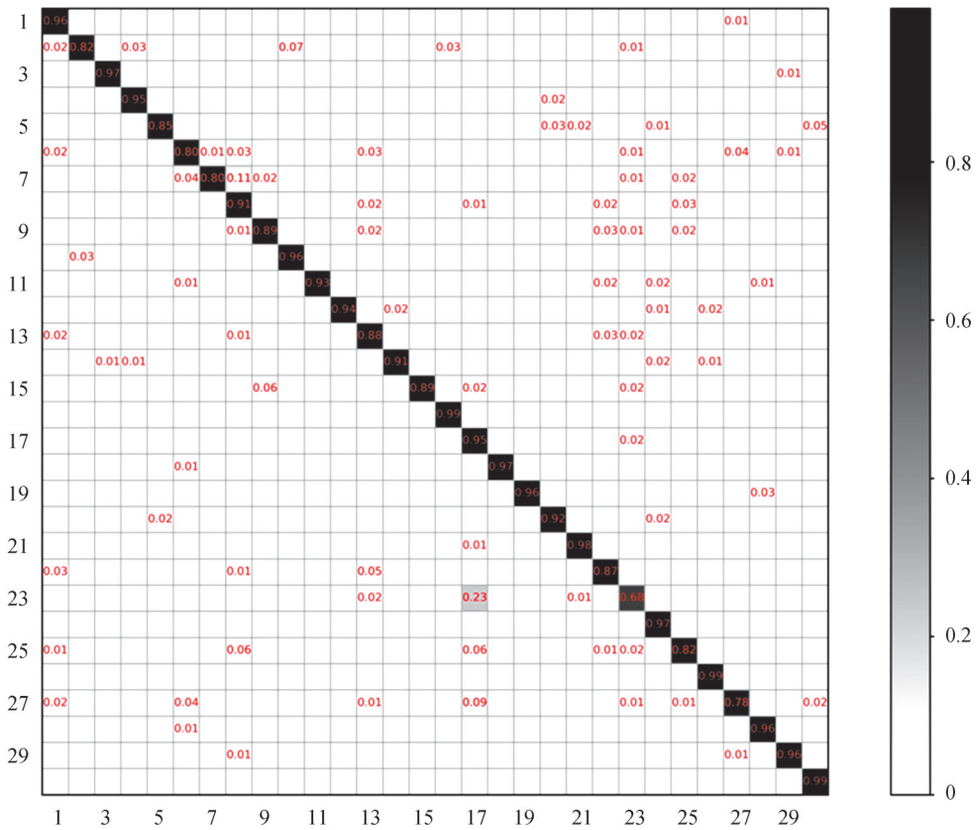

图 7 MCNN-SSAM 在 AID 数据集 20% 训练集比例下的场景分类混淆矩阵

Fig.7 Confusion matrix of scene classification on AID dataset under 20% training ratios

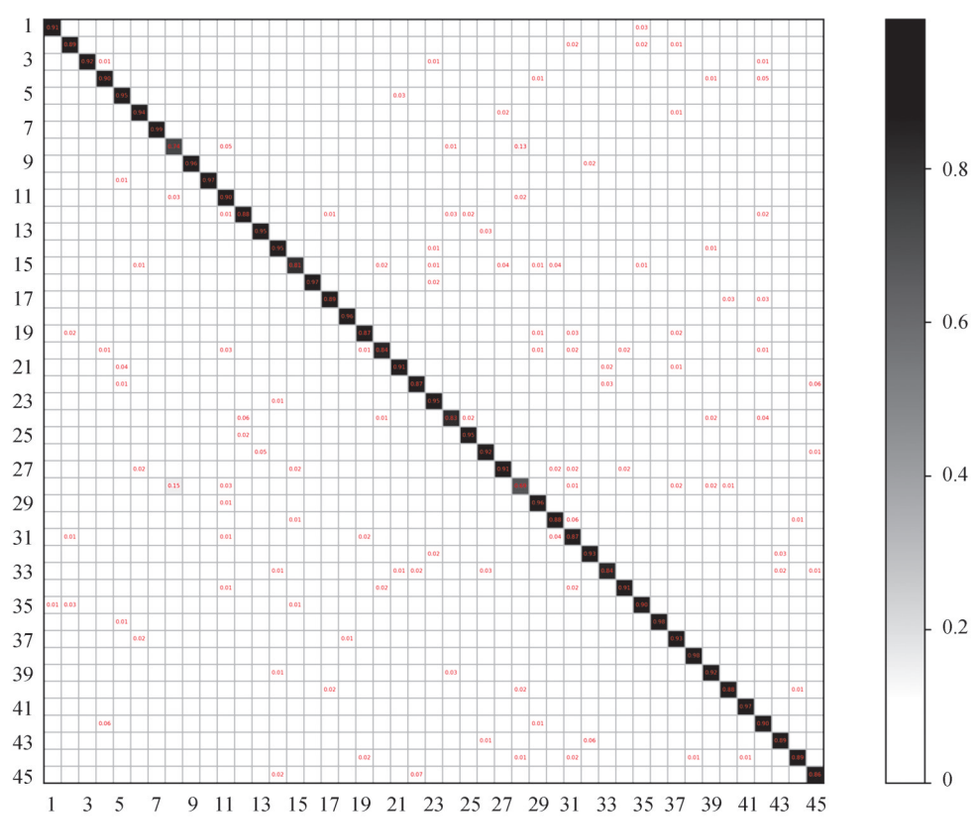

图 8 MCNN-SSAM 在 NWPU45 数据集 20% 训练集比例下的场景分类混淆矩阵

Fig.8 Confusion matrix of scene classification on NWPU45 dataset under 20% training ratios

2.5 可视化实验

本节通过类别激活映射(Class Activation Mappings，CAMs)进行特征图可视化[28]。CAMs通过热图的形式对特征图进行高亮显示，以求从直观效果上可视化CNNs所学习到的特征表述。图9给出MCNN-SSAM与其他相关网络的热图结果对比。

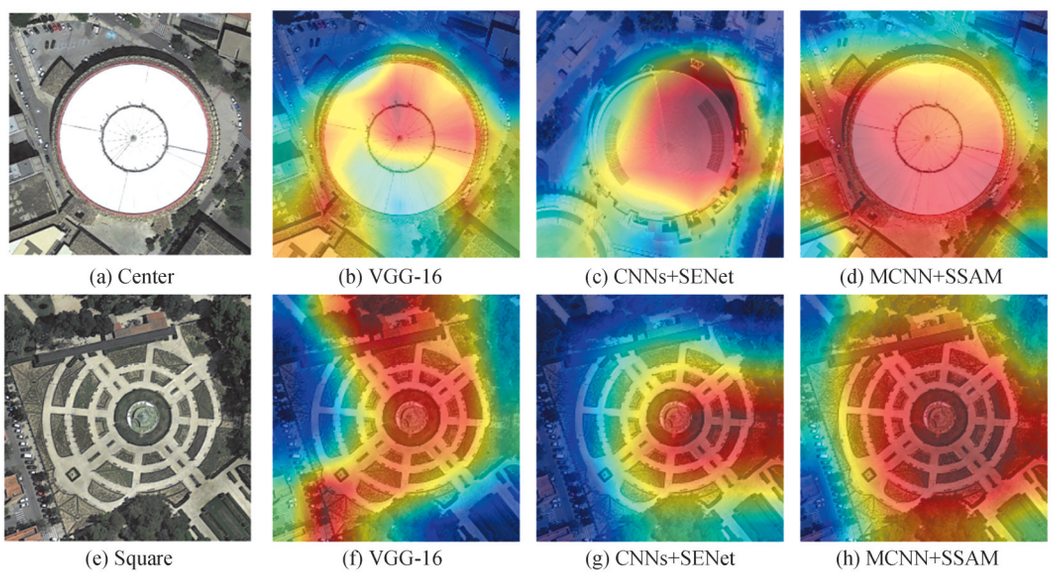

图9 MCNN-SSAM与其他相关网络的热图结果对比

Fig.9 Visual comparison of heatmaps between MCNN-SSAM and other related networks

从图9可以看出，VGG-16所得到的热图虽然能够较准确地聚焦所识别场景区域，但较为发散。例如：在中央广场场景中，VGG-16网络可以捕捉到中央广场所在区域，但范围较小，且不够精确；CNNs+SENet可以更好地将特征识别区域聚集在场景区域，但范围受限。例如：无论是在中央广场场景还是广场场景中，CNNs+SENet网络均能够捕捉到广场区域，且较为集中，但忽略了场景周边环境信息对分类效果的影响；本文所提出的MCNN+SSAM不仅能够将特征识别区域聚集在待识别的场景区域，还可以顾及到周边环境区域，激活区域范围较大，且定位较为准确，故MCNN+SSAM可以得到更优的遥感图像场景分类效果。

2.6 消融实验

MCNN-SSAM中MCNN和SSAM两个模块，为验证这两个模块的有效性，表2给出相关消融实验结果。同时，本小节还将MCNN-SSAM中的稀疏二阶注意力模块替换为基于一阶特征统计量的SENet和基于二阶统计量的协方差注意力机制CovNet，以此验证所提稀疏二阶注意力模块的有效性。

表2 消融和泛化实验结果

Table 2 Results of ablation and generalization experiments

	AID(50%)	NWPU45(20%)
CNNs+SENet	91.01	83.30
CNNs+CovNet	92.97	84.90
CNNs+SSAM	93.36	85.22
MCNN	90.72	83.21
MCNN+SENet	93.34	87.04
MCNN+CovNet	93.97	88.97
MCNN−SSAM	**94.98**	**90.61**

从表2可以看出，无论是CNNs深度特征还是多尺度MCNN深度特征，与基于一阶特征统计量的注意力机制(CNNs+SENet和MCNN+SENet)相比，基于二阶特征统计量的注意力机制(CNNs+CovNet、

CNNs+SSAM、MCNN+CovNet和MCNN+SSAM)所得到的场景分类准确率均有不同程度的提高。同时,通过对比实验可以看出,MCNN模块、SSAM模块以及这两种模块的融合均对遥感图像场景分类效果的提升有促进作用,例如:与CNNs+SSAM相比,MCNN-SSAM的场景分类效果提升了1.62%~5.39%;与CNNs+CovNet相比,MCNN+CovNet的场景分类准确率有1.0%~4.07%的提升。另外,在CNNs实验部分,与CNNs+SENet和CNNs+CovNet相比,CNNs+SSAM有着1.92%~2.35%和0.32%~0.39%的性能提升;在MCCN实验部分,与MCNN+SENet和MCNN+CovNet相比,MCNN-SSAM的场景分类准确率提升了1.64%~3.57%和1.01%~1.64%。验证了本文所提稀疏二阶注意力模块的有效性。

3 结论

针对地面目标尺度信息不同及场景特征描述困难所导致场景分类算法分类效果欠佳的现象,本文从深度特征学习的角度出发,提出稀疏二阶注意力机制驱动的多尺度卷积遥感图像场景分类网络。与单尺度卷积模块相比,金字塔卷积模块所得到的特征图感受野不同,且能够增强深度特征表述能力;稀疏二阶注意力模块利用稀疏二阶统计量进行通道相关性的计算,在顾及二阶特征统计量特征冗余性的同时,达到了特征增强效果。从场景分类精确性、混淆矩阵、可视化等多个方面的对比实验表明:本文所提MCNN-SSAM在两个具有挑战性的遥感图像场景分类数据集上有较好的遥感图像场景分类效果。但本文所提网络需要手动设置的参数量较多,在未来的工作中,如何设计超参数较少且可生成鲁棒稀疏二阶统计量的算法,并将其嵌入至网络中进行端到端训练将是未来的工作重点之一;另外,构建一景遥感场景图像分类数据集,并在一景典型遥感图像中验证本文所提网络的有效性同样值得我们关注。

参考文献

[1] CHENG Gong, YANG Ceyuan, YAO Xiwen, et al. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(5): 2811-2821.

[2] SHI Cuiping, WANG Peng, WANG Liguo. Branch feature fusion convolution network for remote sensing scene classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 5194-5210.

[3] PENG Cheng, LI Yangyang, JIAO Licheng, et al. Efficient convolutional neural architecture search for remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(7): 6092-6105.

[4] ZHOU Peicheng, CHENG Gong, YAO Xiwen, et al. Machine learning paradigms in high-resolution remote sensing image interpretation[J]. National Remote Sensing Bulletin, 2021, 25(1):182-197.
周培诚, 程塨, 姚西文, 等. 高分辨率遥感影像解译中的机器学习范式[J]. 遥感学报, 2021, 25(1): 182-197.

[5] TAO Chao, QI Ji, LU Weipeng, et al. Remote sensing image scene classification with self-supervised paradigm under limited labeled samples[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 8004005.

[6] WANG Jianan, GAO Yue, SHI Jun, et al. Scene classification of optical high-resolution remote sensing images using vision transformer and graph convolutional network[J]. Acta Photonica Sinica, 2021, 50(11): 1128002.
王嘉楠, 高越, 史骏, 等. 基于视觉转换器和图卷积网络的光学遥感场景分类[J]. 光子学报, 2021, 50(11): 1128002.

[7] NI Kang, LIU Pengfei, WANG Peng. Compact global-local convolutional network with multifeature fusion and learning for scene classification in synthetic aperture radar imagery [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 7284-7296.

[8] XIE Jie, HE Nanjun, FANG Leyuan, et al. Scale-free convolutional neural network for remote sensing scene classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(9): 6916-6928.

[9] CHENG Gong, LI Zhenpeng, YAO Xiwen, et al. Remote sensing image scene classification using bag of convolutional features[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(10): 1735-1739.

[10] HE Nanjun, FANG Leyuan, LI Shutao, et al. Remote sensing scene classification using multilayer stacked covariance pooling[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(12): 6899-6910.

[11] NI Kang, WU Yiquan. Scene classification from remote sensing images using mid-level deep feature learning [J]. International Journal of Remote Sensing, 2020, 41(4): 1415-1436.

[12] YANG Zhou, MU Xiaodong, WANG Shuyang, et al. Scene classification of remote sensing images based on multiscale features fusion[J]. Optics and Precision Engineering, 2018, 26(12): 3099-3107.
杨州, 慕晓冬, 王舒洋, 等. 基于多尺度特征融合的遥感图像场景分类[J]. 光学精密工程, 2018, 26(12): 3099-3107.

[13] LU Xiaoqiang, SUN Hao, ZHENG Xiangtao. A feature aggregation convolutional neural network for remote sensing

scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(10):7894-7906.
[14] LI Ruoyao, ZHANG Bo, WANG Bin. Remote sensing image scene classification based on multilayer feature context encoding network[J]. Journal of Infrared and Millimeter Waves, 2021, 40(4): 530-538.
李若瑶, 张铂, 王斌. 基于多层特征上下文编码网络的遥感图像场景分类[J]. 红外与毫米波学报, 2021, 40(4): 530-538.
[15] HU Jie, SHEN Li SAMUEL A, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.
[16] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision, 2018:3-19.
[17] CAO Yue, XU Jiarui, LIN Stephen, et al. Gcnet: non-local networks meet squeeze-excitation networks and beyond[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019: 1971-1980.
[18] WANG Qilong, WU Banggu, ZHU Pengfei, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 11531-11539.
[19] GAO Zilin, XIE Jiangtao, WANG Qilong, et al. Global second-order pooling convolutional networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 3024-3033.
[20] BRYAN B, GONG Yuan, ZHANG Yizhe, et al. Second-order non-local attention networks for person re-identification [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 3760-3769.
[21] DUTA I C, LIU Li, ZHU Fan, et al. Pyramidal convolution: rethinking convolutional neural networks for visual recognition [EB/OL].[2020-06-20].https://arxiv.org/abs/2006.11538.
[22] GAO Zhi, WU Yuwen, ZHANG Xiaoxun, et al. Revisiting bilinear pooling: a coding perspective[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 3954-3961.
[23] LI Xiao, LEI Lin, KUANG Gangyang. Locality-constrained bilinear network for land cover classification using heterogeneous images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 2501305.
[24] XIA Guisong, HU Jingwen, HU Fan, et al. AID: a benchmark data set for performance evaluation of aerial scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.
[25] CHENG Gong, HAN Junwei, LU Xiaoqiang. Remote sensing image scene classification: benchmark and state of the art [J]. Proceedings of the IEEE, 2017, 105(10): 1865-1883.
[26] CAO Ran, FANG Leyuan, LU Ting, et al. Self-attention-based deep feature fusion for remote sensing scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 18 (1): 43-47.
[27] ZHANG Wei, TANG Ping, ZHAO Lijun. Remote sensing image scene classification using cnn-capsnet[J]. Remote Sensing, 2019, 11 (5): 494.
[28] SELVARAJU R R, COGWELL M, DAS A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128 (2): 336-359.

Multi-scale Convolutional Neural Network Driven by Sparse Second-order Attention Mechanism for Remote Sensing Scene Classification

NI Kang[1,2], ZHAO Yuqing[3], CHEN Zhi[1]
(1 *School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing* 210023, *China*)
(2 *Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing* 210023, *China*)
(3 *School of Management and Engineering Capital University of Economics and Business, Beijing* 100070, *China*)

Abstract: Remote sensing image scene classification is one of the important research contents of remote sensing image interpretation. Nowadays, with the rapid development of satellite imaging techniques, remote sensing scene classification which uses High Spatial Resolution (HSR) remote sensing images has received, considerable attention recently, as can be used in natural hazards detection, traffic control, and object detection etc. Based on the feature representation used for remote sensing scene classification, the existing scene classification approaches can be categorized into three classes: handcrafted feature based methods, unsupervised feature learning-based methods, and deep feature learning-based methods. Convolutional Neural Networks (CNNs), one of the deep feature learning-based methods, have achieved great success in the computer vision community. Especially, the powerful feature representations learned through CNNs have been widely used in remote sensing scene classification, but due to the different scale

information of ground targets and the complex spatial distribution and texture information of the scene images, the classification effect of the scene classification algorithm based on CNN is insufficient good. For addressing the above problems, the paper proposes a multi-scale convolutional neural network driven by a sparse second-order attention mechanism (MCNN-SSAM) while comprehensively considering the accuracy of scene classification and feature dimensions. The proposed MCNN-SSAM network includes the following parts: backbone network, pyramid convolution module, sparse second-order attention module and softmax classification layer. The network firstly inserts a multi-scale convolution layer after the backbone network to acquire the characteristic expressions of different scale information targets of the ground target, and embeds the group convolution into the multi-scale convolution layer to reduce the computational complexity; Secondly, after discuss the advantage of the attention mechanism of first and second-order statistics, a sparse second order attention mechanism is proposed to enhance the discriminability of channel information of different scale convolution features. The sparsity of the attention mechanism is able to effectively reduce the feature dimension of the second-order statistics while ensuring the performance of scene classification; Finally, the multi-scale convolutional layer and the sparse second-order attention mechanism are embedded into the proposed network for end-to-end training. We conduct extensive experiments on two challenging high-resolution remote sensing data sets, i.e., AID (Aerial Image Dataset) and NWPU45 (NWPU-RESISC45) datasets. The AID dataset contains 10 000 images in RGB space, which has 30 different scene classes and of size 600×600 in each class; There are 31 500 optical RS images for 45 scene classes, and each image measures 256×256 pixels on the NWPU45 dataset. In this paper, the VGG-16 network is selected as the backbone of MCNN-SSAM, and the Adam optimizer is used for end-to-end training. The training parameters of the proposed network are set as follows: initial learning rate 0.001, weight attenuation coefficient 0.001, batch size 32, momentum 0.9. All experiments are implemented in PyTorch, NVIDIA GeForce GTX 8G 1070 Ti GPU, and 32.00 GB RAM. we make the experimental result on AID dataset to analyze the influence of some important parameters on the MCNN-SSAM, then we can conclude that the number of the atoms in the dictionary and low-rank matrix parameters have a greater impact on the remote sensing scene classification performance of the proposed MCNN-SSAM. Afterwards, we compare MCNN-SSAM with some related networks, i.e., AlexNet, VGG-16, SAFF, MSCP, and CapsNet. The experimental results show that: compared with the benchmark network (VGG-16), the overall accuracy (OA) of MCNN-SSAM is improved by 5.27%～5.34% and 10.20%～10.82%; While compared with the SAFF, MSCP, and CapsNet networks, the remote sensing scene classification accuracy is improved by 0.23%～1.61% and 1.34%～2.75%. Additionally, based on the confusion matrix, we can observe that most of the remote sensing scene classes can be classified easily and correctly, some even achieving high classification accuracies, i.e., mountains and viaducts in the AID dataset, jungles and sea ice in the NWPU45 dataset. Meanwhile, the effectiveness of the Sparse Second-order Attention Mechanism (SSAM) is verified by comparing with other related attention mechanisms and heat map results. Finally, we make the ablation generalization experiments to verify the effectiveness of MCNN-SSAM, such as SENet (Squeeze-and-Excitation Networks), CovNet which is based on covariance statistics, and SSAM. We can conclude that, whether the CNN features or multi-scale MCNN features, compared with the attention mechanism based on first-order feature statistics (CNN+SENet and MCNN+SENet), The scene classification accuracy obtained by CNN+CovNet, CNN+SSAM, MCNN+CovNet, and MCNN+SSAM which are based on the attention mechanism of second-order feature statistics has been further improved. In addition, MCNN module, SSAM module, and the fusion of these two modules can improve the classification accuracy of remote sensing image scene images. In this paper, we propose a multi-scale convolutional neural network driven by sparse second-order attention mechanism for remote sensing scene classification. The experiment results illustrate that the proposed MCNN-SSAM improves the accuracy of remote sensing image scene classification while taking into the feature dimensions of the second order feature statistics.

Key words: Remote sensing image; Convolutional neural networks; Feature distinguishability; Multi-scale convolution; Attention mechanism

OCIS Codes: 100.4996; 100.1830; 150.0155; 100.3008

Foundation item: National Natural Science Foundation of China (No. 62101280), Natural Science Foundation of Jiangsu Province (No. BK20210588), Nanjing University of Posts and Telecommunications Science Foundation (No. NY220135)