文章编号:1001-9014(2021)04-0530-09

DOI:10.11972/j.issn.1001-9014.2021.04.012

基于多层特征上下文编码网络的遥感图像场景分类

李若瑶[1,2], 张 铂[1,2], 王 斌[1,2*]

(1. 复旦大学 电磁波信息科学教育部重点实验室,上海 200433;
2. 复旦大学 信息学院智慧网络与系统研究中心,上海 200433)

摘要:遥感图像场景分类问题是目前遥感图像处理领域中的研究热点之一。卷积神经网络(Convolutional Neural Networks, CNNs)具有强的特征提取能力,已被广泛应用于遥感图像场景分类中。然而,目前的方法并没有充分考虑并利用CNN不同层间的互补信息和遥感图像的空间上下文信息,导致其相应的分类精度有待提高。针对上述问题,提议一种多层特征上下文编码网络,并将其用于解决遥感图像场景分类问题。所提议网络由两部分组成:1)密集连接的主干网络;2)多尺度上下文编码模块。前者用于融合CNN不同层的特征信息,后者用于对蕴含在多层特征中的空间上下文信息进行编码利用。在两个大规模遥感图像数据集上的实验结果表明,与现有的遥感图像场景分类方法相比,所提出的网络框架取得了显著的分类精度提升。

关 键 词:遥感图像;场景分类;卷积神经网络;多层特征上下文编码;空间上下文信息

中图分类号:TP751 **文献标识码**:A

Remote sensing image scene classification based on multilayer feature context encoding network

LI Ruo-Yao[1,2], ZHANG Bo[1,2], WANG Bin[1,2*]

(1. Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China;
2. Research Center of Smart Networks and Systems, School of Information Science and Technology, Fudan University, Shanghai 200433, China)

Abstract: Remote sensing image scene classification is one of the current hot topics in the field of remote sensing image processing. Since convolutional neural networks (CNNs) have powerful feature extraction capabilities, they have been widely applied in remote sensing image scene classification. However, the current methods have not fully considered and utilized the complementary information between different layers of CNN and the spatial context information of remote sensing images, resulting in that the corresponding classification accuracy needs to be improved. In order to address these issues, a multilayer feature context encoding (MFCE) network is proposed and utilized to solve the problem of scene classification for remote sensing images. The proposed network is composed of two parts: 1) A densely connected backbone; 2) A multiscale context encoding (MCE) module. The former is adopted to fuse the feature information of different layers of CNN, and the latter is utilized to encode and exploit the spatial context information that resides in the multilayer features. Experimental results on two large-scale remote sensing image datasets demonstrate that compared with the existing remote sensing image scene classification methods, the proposed network framework can achieve a significant gain in classification accuracy.

Key words: remote sensing images, scene classification, convolutional neural network (CNN), multilayer feature context encoding (MFCE), spatial context information

收稿日期:2020-10-09,**修回日期**:2021-04-21 **Received date**:2020-10-09,**Revised date**:2021-04-21

基金项目:国家自然科学基金(61971141, 61731021)

Foundation items:Supported by National Natural Science Foundation of China (61971141 and 61731021)

作者简介(Biography):李若瑶 (1995-),女,天津人,硕士研究生,主要研究领域为遥感图像场景分类

E-mail: 18210720036@fudan.edu.cn

*通讯作者(Corresponding author): E-mail: wangbin@fudan.edu.cn

引言

随着遥感和卫星技术不断进步,大量的高分辨率遥感图像被获取,并用于地表信息分析[1]。遥感图像场景分类旨在根据一幅高分辨率遥感图像的内容为其赋予相应的语义标签,可用于城市制图,环境监测等实际应用中,是当前的研究热点问题。然而,高分辨率遥感图像具有复杂的纹理信息和空间分布,通常具有类内差异大而类间差异小的特点,这一固有特点给遥感图像场景分类任务造成了较大的困难。

遥感图像场景分类通常在特征空间中进行,在过去的几十年中,大量工作致力于为高分辨率遥感图像构建鲁棒的特征表示。早期的遥感图像场景分类工作主要使用了手工设计的特征,例如尺度不变特征变换(Scale-Invariant Feature Transform, SIFT)[2],梯度方向直方图(Histogram of Oriented Gradients, HOG)[3]等。但是,由于手工设计特征往往不能够充分提取遥感图像中丰富的语义信息,其方法的分类性能难以满足实际应用的需求。随着机器学习和人工智能的快速发展,卷积神经网络(Convolutional Neural Network, CNN)在不同的计算机视觉任务中展现了优势。由于CNN强大的特征表征能力,许多遥感图像场景分类工作开始使用CNN作为特征提取器,并取得了较为显著的性能提升。现阶段,基于CNN的遥感图像场景分类方法大致可分为两类:基于单层特征的方法和基于多层特征的方法。

第一类方法将现有的用于自然图像分类的CNN模型,如AlexNet[4]或VGGNet[5],迁移应用到遥感图像场景分类领域,将它们作为特征提取器,使用全连接层的特征对遥感图像进行分类。先前的工作[6]直接使用通用的预训练CNN模型对遥感图像进行分类。在先前工作的基础之上,具有判别力的卷积神经网络模型(Discriminative-CNN, D-CNN)[7],通过使用度量学习正则项,进一步增强了现有网络模型的判别力。在上述方法中,它们只利用了全连接层的特征,却忽视了来自CNN卷积层的特征。然而,根据已有的研究工作[8-10],CNN中不同层的特征包含着不同的空间信息和语义信息,浅层特征包含有更多的空间结构信息,深层特征则包含更为丰富的语义信息。由于高分辨率遥感图像具有较为复杂的空间结构信息,我们考虑,将不同层的特征进行融合,以充分利用多层特征中的空间信息和语义信息,这些将对遥感图像场景分类任务具有较为重要的积极作用。

针对上面提出的问题,第二类方法融合CNN不同层的特征,然后使用不同的特征编码方法增强特征表征能力。多层堆叠协方差池化(Multilayer Stacked Covariance Pooling, MSCP)[11]融合了预训练的CNN模型中来自不同卷积层的特征并使用协方差池化方法获取多层特征的二阶信息。特征聚合卷积神经网络(Feature Aggregation Convolutional Neural Network, FACNN)[12],是一个端到端的网络结构,可在训练过程中通过遥感图像的标签信息有监督地融合网络的多层特征,同时对多层特征进行编码。尽管上述方法能够提升遥感图像场景分类任务的结果,这些方法仍然存在以下两个主要不足:首先,上述利用多层特征的方法,手工选取来自特定层的特征,并进行融合。这样的做法一方面只能利用特定层的特征,其余层的被忽略;另一方面则需要进行大量实验比较融合不同层的分类结果,以选取合适的层的特征进行融合,极为耗时。其次,尽管空间上下文信息在遥感图像场景理解中至关重要,这些方法却未将空间上下文信息纳入考虑。一幅高分辨率遥感图像中往往包含多种来自不同区域和不同尺度的地面覆盖物单元[12]。地面覆盖物单元的空间分布很大程度上决定了一幅遥感图像的场景类别。例如,判断一幅场景是属于铁路或是火车站,需要通过站台建筑和其周围区域铁路的组合信息而决定,如果只利用铁路的信息,很容易将火车站类别的场景错误分类为铁路。因此,如果不能充分利用空间上下文信息,将会对遥感图像场景分类结果造成负面影响。

为解决上述问题,本文设计一个多层特征上下文编码(Multilayer Feature Context Encoding, MFCE)网络,在融合多层特征的同时,对空间上下文信息进行编码利用。该网络由两个部分组成:1)进行多层特征提取的主干网络,产生融合空间结构信息和语义信息的特征表示;2)多尺度上下文编码(Multi-scale Context Encoding, MCE)模块,用于充分挖掘蕴含在多层特征中的多尺度空间上下文信息。为设计一个可行的用于遥感图像场景分类的网络框架,本文的主干网络采用了用于自然图像分类的稠密网络(DenseNet)[13]。该网络通过促进浅层到深层特征复用,在自然图像分类任务中取得了令人满意的结果,同时具有参数量小、抗过拟合的优良性质。

该网络结构无需手工选取需要进行融合的特征，可以通过端到端训练的方式自动学习不同层特征的信息互补关系，实现深层特征和浅层特征的融合。这里，基于高分辨率遥感图像空间结构信息复杂的特点，我们将其引入遥感图像场景分类任务中。同时，为进一步获取多层特征中的空间信息，本文设计了一个MCE模块，该模块包括多尺度池化和上下文编码两个部分。该模块对深度网络提取的特征进行多尺度池化，以获取图像不同子区域和不同尺度的信息，并采用上采样操作，以保持不同尺度特征间的空间位置关系；与此同时，更进一步增加对多尺度特征图的端到端编码过程，实现对不同子区域的空间上下文关系的学习，从而提升分类精度。在两个大规模遥感数据集上的实验结果表明，本文所提出的网络框架可有效提升遥感图像场景分类的精度。

1 相关工作

本节通过介绍DenseNet的详细结构信息，以进一步阐释其融合空间结构信息与语义信息的机制[13]。如图1所示，DenseNet由多个串联而成的密集块（Dense Block）构成。每两个密集块中间为一过渡层（Transition Layer），用以进行空间维度的下采样和通道维度的裁剪。图1的上半部分展示了一个密集块的内部结构，一个密集块中包含若干卷积层，假设密集块的输入为x_0，密集块中的第l层会接收到其之前所有层的输出特征图，第l层的输出可以通过式(1)进行计算：

$$x_l = H_l\left[x_0, x_1, \cdots, x_{l-1}\right] \quad , \quad (1)$$

其中，H_l代表批标准化（Batch Normalization，BN）[14]、线性整流单元（Rectified Linear Unit，ReLU）和核尺寸为3×3的卷积运算组成的复合函数，$\left[x_0, x_1, \cdots, x_{l-1}\right]$代表第$0, 1, \cdots, l-1$卷积层的输出特征图在通道维度上的拼接。

DenseNet的密集块中采用的特征连接方式使得同一个密集块中的每一卷积层能够接收到其之前所有卷积层的输出，因此来自网络浅层的特征可以被网络的深层复用，达到浅层和深层特征融合的效果。研究表明[8-9]，CNN的浅层特征包含更多的空间结构信息，而深层特征包含更多的语义信息，将不同层的特征进行融合，可以更加充分地获取不同层间的互补信息，从而提高网络的分类精度。此外，由于密集块中每一层都通过拼接的方式与前级进行连接，则每一层都参与到特征融合的过程中，网络通过训练自动调整不同层特征在最终分类决策中的权重，从而省去了手工选取某些层特征进行融合所需要的大量参数实验。

此外，DenseNet在进行网络设计时，将密集块中各个卷积层输出的特征图的通道数设定为k，k在DenseNet中被称为生长率（Growth Rate），这是一个预先设定好的超参数。DenseNet的实验结果表明[13]，使用较小的k（例如12），就可以得到很好的分类性能。由于采用了上述设计方式，DenseNet具有参数量小，易于训练，抗过拟合的优良特性。

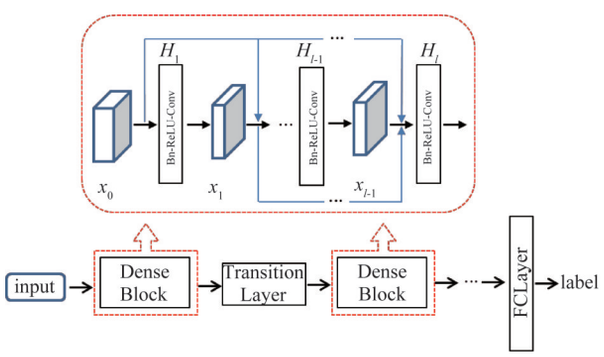

图1 DenseNet的网络结构示意图

Fig. 1 The illustration of the architecture of DenseNet

2 模型构建

本文旨在设计一种同时考虑空间上下文信息和多层特征融合的网络框架。该网络的整体结构如图2所示，由两部分组成：密集连接的主干网络和MCE模块。首先，我们对MFCE网络中采用的主干网络进行介绍，然后，对MCE模块的详细结构信息进行描述，最后，给出了MFCE网络整体训练使用的优化目标函数。

2.1 主干网络

由于CNN强大的特征提取与表征能力，多种CNN结构都曾被用作遥感图像场景分类任务的特征提取器[1]。先前基于神经网络的工作使用了CNN全连接层的特征对遥感图像进行分类[6][7]。而进一步的研究表明[11-12]，由于高分辨率遥感图像空间结构信息较为复杂，将具有丰富空间结构信息的浅层特征和富含语义信息的深层特征进行融合，可以提高神经网络的特征表达能力，进而提高分类精度。然而，现存的使用神经网络多层特征的方法主要通过手工选择的方式对来自CNN特定层的特征进行融合[11-12]，这样的方式不仅需要大量的参数实验以

确定需要进行特征融合的层,另一方面,网络中只有特定层的特征得到了利用,而其它层的特征往往被忽略。

针对上述问题,本文采用目前为止用于自然图像分类的DenseNet作为我们所提议的MFCE网络的主干网络。该网络采用密集连接的机制,实现了浅层和深层的特征的充分融合,同时可通过训练,自动调整不同层特征在最终分类决策中的权重,从而省去了手工选取特定层特征进行融合所需的大量参数实验[13]。在本文所提出的MFCE网络中,移除了DenseNet-121的全连接层,只将其剩余部分作为MFCE网络的主干网络,以实现多层特征融合的目的。

2.2　多尺度上下文编码模块

为进一步挖掘蕴含在CNN多层特征中的空间上下文信息,本文提出了一个MCE模块,其具体结构如图2所示。该模块由两部分组成:多尺度池化和上下文编码。

多尺度池化旨在对主干网络提取的特征进行不同粒度的采样,获取图像中不同位置、不同尺度的空间信息。粒度越小,即尺度更小的池化,可以反映更多的图像空间细节信息;而大尺度池化,可以反映图像的更大子区域的信息。本文通过多级平均池化操作以获取图像不同位置子区域的多尺度信息。在本文设计中,将主干网络提取的特征记为M,其维度为$H \times W \times N$,其中H和W分别为特征的高度和宽度,N为通道数。对M采用4级多尺度池化,每级池化后的输出特征图的空间维度分别设置为2×2, 4×4, 6×6和8×8,同时在每一个池化层级后面都配置一层1×1卷积层,将多尺度特征图通道数裁剪为原有通道数的1/4,以减少冗余信息。将经过池化和通道裁剪后的特征分别记为m_1, m_2, m_3和m_4,其对应维度分别为$2 \times 2 \times N/4$, $4 \times 4 \times N/4$,$6 \times 6 \times N/4$和$8 \times 8 \times N/4$。

上下文编码用于对来自于不同区域和尺度的特征图的空间关系进行编码。首先,使用双线性插值对m_1, m_2, m_3和m_4进行上采样操作。由于m_1, m_2, m_3和m_4反映了图像中不同区域、不同尺度的空间信息,上采样操作能够维持它们的相对空间位置关系。将经过上采样的m_1, m_2, m_3, m_4与M进行通道维度拼接,则拼接后的特征图C可表示为:

$$C = \left[U\left(m_1\right), U\left(m_2\right), U\left(m_3\right), U\left(m_4\right), M\right], \quad (2)$$

其中,$U(\cdot)$表示上采样运算,$\left[U\left(m_1\right), U\left(m_2\right), U\left(m_3\right), U\left(m_4\right), M\right]$表示上采样后的多尺度特征与原有特征$M$在通道维度上的拼接。拼接后的特征$C$集合了来自图像不同尺度、不同位置的子区域的信息,维度为$H \times W \times 2N$。对拼接后的特征C使用1×1卷积,以对不同特征图的关系进行编码,通过参数学习得到编码后的特征F,其维度为$H \times W \times N$。

2.3　优化目标

为构建MFCE网络,我们将提出的MCE模块连接到密集连接的主干网络之后,然后对MFCE网络所提取的特征进行全局平均池化。通过全局平均池化操作,可获取特征中每个通道的均值,该操作在不同的分类网络中展现出较好的分类性能[15][16]。

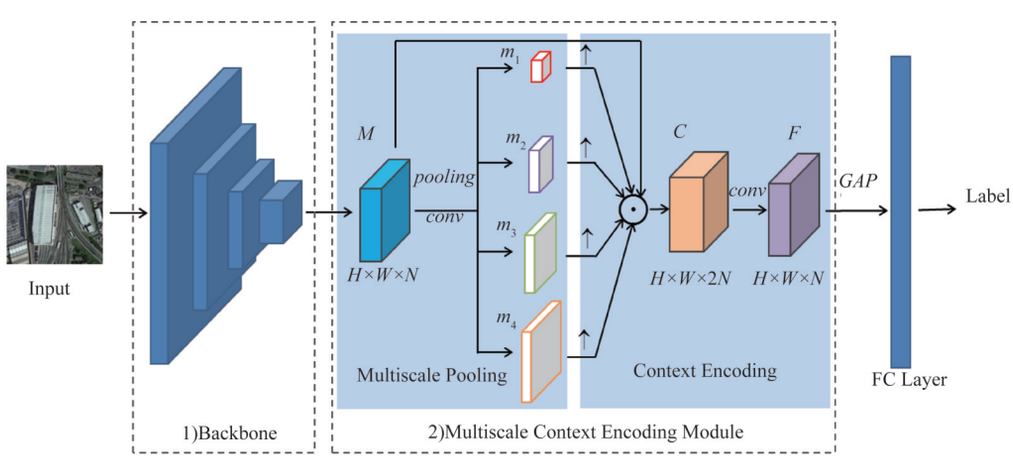

图2　提出的MFCE的网络框架

注:⊙和↑分别表示通道维度的拼接操作和空间维度的上采样操作

Fig. 2　The framework of the proposed MFCE network

Note: ⊙ and ↑ denote the channel concatenation operation and the spatial up-sampling operation, respectively

MFCE的输出特征F具有$H \times W \times N$的维度，则全局平均池化操作可以由下式表示：

$$g_n = \frac{\sum_{w=1}^{W}\sum_{h=1}^{H} F_{w,h,n}}{W \times H} \quad , \quad (3)$$

其中，g_n表示通道n的均值。经过全局平均池化后，将得到的特征输入全连接层以产生最终的分类预测结果。

所提出的MFCE网络通过密集连接的主干网络和MCE模块两阶段处理的方式，实现了神经网络多层信息融合以及空间上下文信息的利用。所提议网络可进行端到端训练，且易于实现。采用最小化交叉熵损失函数对网络进行训练，交叉熵损失函数如下式所示：

$$loss = -\frac{1}{L}\sum_{i=1}^{L}\left\langle \mathbf{y}_i, \log\left(\Phi\left(\mathbf{x}_i\right)\right)\right\rangle \quad , \quad (4)$$

其中，$\mathbf{x}_i$表示训练过程中总共L张图像中的第i个样本的全连接层输出，其对应的类别标签为$\mathbf{y}_i$。$\Phi\left(\mathbf{x}_i\right)$代表softmax函数，而$\left\langle \mathbf{y}_i, \Phi\left(\mathbf{x}_i\right)\right\rangle$则代表标签$\mathbf{y}_i$和$\Phi\left(\mathbf{x}_i\right)$的内积运算。

3 实验结果与分析

本节在两个大规模数据集上对所提出的MFCE网络进行性能评估。首先，对数据集和实验设置的具体细节进行介绍，然后，将MFCE网络与现存的遥感图像场景分类方法的分类精度进行对比，最后，对所提出的MFCE网络各部分作用及相关参数设置进行了详尽分析，还给出了网络所需的参数量和计算量。

3.1 实验数据

本文采用Aerial Image dataset（AID）①和NWPU-RESISC45②两个大规模数据集对所提出方法的性能进行评估[17-18]。AID数据集包含10 000张采集自Google Earth的尺寸为600×600像素的图像，空间分辨率从8 m到0.5 m不等，每类图像的数目变化范围从220到420不等。NWPU-RESISC45数据集包括45类采集自Google Earth的共31500张遥感图像，每幅图像尺寸为256×256像素，空间分辨率从30 m到0.2 m不等。上述两个数据集中的图像采集于全世界范围内的不同国家和地区，拍摄于不同的成像平台以及气候条件下。同时，这两个数据集也是迄今为止遥感图像场景分类领域中规模最大的数据集，含有丰富的场景类别。因此，在AID数据集和NWPU-RESISC45数据集上，对遥感图像场景分类方法进行测试，所得结果可以较好地反映该方法在实际任务中的性能，对于验证方法的普适性具有较大的参考意义。

上述两个数据集中的图像同时具有空间布局复杂、纹理信息丰富的特点。一幅图像中往往包含多种来自不同区域和不同尺度的地面覆盖物单元，造成不同类图像间的相似度较大，场景类别通常取决于地面覆盖物的空间布局，分类具有较大难度。图3给出了上述两个数据集的样例图像。如图3(a)所示，操场和体育馆类别中均包含运动场，两类图像的主要区别在于周围是否有看台；在图3(b)中，火车站和铁路类别中均包含铁路，两类图像的主要区别在于周围是否有站台建筑。

3.2 评价指标和实验设置

实验中，采用总体精度（Overall Accuracy，OA），即被正确分类的图像数量除以图像总数，作为实验的评价指标，以量化评估分类结果的精度。

对于AID数据集，根据已有工作[7,11-12]的数据集划分方式，在每类图像中分别随机选50%和20%的图像作为训练集，其余作为测试集。对于NWPU-

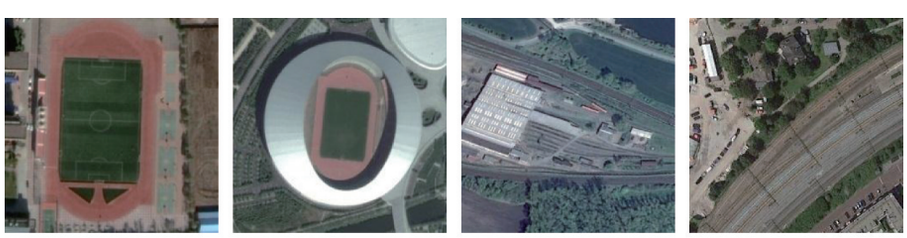

图3 遥感图像样例图像 (a) AID数据集，(b) NWPU-RESISC45数据集

Fig. 3 Samples of remote sensing images (a) AID dataset, (b) NWPU-RESISC45 dataset

①https://captain-whu.github.io/AID/

②http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html

RESISC45数据集,遵循文献[7,11]中的实验设置,在每类图像中分别随机选取10%和20%的图像作为训练集,其余90%和80%作为测试集。所有的实验均重复5次,以提供更为可靠的结果。

本文提出的模型在Pytorch框架上进行构建,并使用NVIDIA TITAN XP的单个GPU对模型进行训练,该GPU具有12 GB的内存。MFCE网络初始学习速率设置为0.005,采用权值衰减为10^{-4}和动量为0.9的随机梯度下降(Stochastic Gradient Descent, SGD)方法对网络参数进行优化,训练每进行10个epoch,学习速率调整为原来的1/10。在训练进行到设定的最大epoch数时(本文设置为25),训练终止。在训练时对数据集进行数据增广,对每张图像进行随机镜像和旋转。

3.3　与其他方法的对比

为验证所提出网络框架的分类精度,在两个大规模数据集上,将所提出的MFCE网络的分类结果与其它遥感图像场景分类方法进行对比。对比方法均为基于CNN的深度模型,包括两种基线模型(VGGNet-16、DenseNet-121)和几种现存的用于遥感图像场景分类的最优方法D-CNN[7], MSCP[11]和FACNN[12]。其中,D-CNN是使用CNN单层特征的方法,该方法通过引入度量学习正则项,增加CNN网络的判别力,显著提升了不同网络结构的分类精度。而MSCP和FACNN则是使用CNN多层特征的方法。MSCP使用预训练的CNN作为特征提取器,然后将提取出的多层特征进行堆叠,使用协方差池化方法充分挖掘CNN的多层特征中的互补信息,其分类精度相较于基线方法取得了较大提高。FACNN在训练过程中,将网络不同层的特征进行聚合拼接,然后对聚合后的特征在图像标签信息的监督下进行编码,其在不同数据集上都取得了较高的分类精度。

表1和表2展示了在两个大规模数据集上不同方法的分类精度。通过比较MFCE, MSCP, FACNN和Fine-tuned DenseNet-121的分类精度,可以发现,MFCE网络在分类精度上实现了较大幅度的提高,这表明空间上下文信息的利用对提升遥感图像场景分类的精度至关重要。而这些方法与D-CNN的对比结果则表明,融合网络不同层的特征,将空间结构信息与语义信息进行融合,对遥感图像场景分类问题有积极作用。综合以上,MFCE的优势归功于两个方面:首先,MCE模块可以充分利用遥感图像的上下文信息;其次,MFCE采用了密集连接的机制,可以有效获取包含多层特征信息的鲁棒特征表示,从而增强了网络的特征表示能力。

表1　AID数据集上不同训练集比例下各种方法的分类精度

Table 1　OA of different methods on AID dataset with different training ratios

Method	OA	
	Tr=20%	Tr=50%
VGG-VD-16[17]	86.59±0.29	89.64±0.36
Fine-tuned DenseNet-121	94.75±0.18	96.56±0.17
FACNN[12]	-	95.45±0.11
D-CNN with VGGNet-16[7]	90.82±0.16	96.89±0.10
VGG-VD16+MSCP+MRA[11]	92.21±0.17	96.56±0.18
MFCE (ours)	**95.51±0.09**	**97.14±0.19**

表2　NWPU-RESISC45数据集上不同训练集比例下各种方法的分类精度

Table 2　OA of different methods on NWPU-RESISC45 dataset with different training ratios

Method	OA	
	Tr=10%	Tr=20%
VGGNet-16[18]	87.15±0.45	90.36±0.18
Fine-tuned DenseNet-121	91.56±0.21	93.72±0.20
FACNN[12]	-	-
VGG-VD16+MSCP+MRA[11]	88.07±0.18	90.81±0.13
D-CNN with VGGNet-16[7]	89.22±0.50	91.89±0.22
MFCE (ours)	**92.42±0.20**	**94.40±0.09**

3.4　分析

3.4.1　不同池化等级对网络分类效果的影响

在所提出的MCE模块中,采取了多尺度池化运算,以获取图像不同粒度的特征图,反映了图像不同尺度的特征。因此,本节分别考察不同池化等级对分类结果的影响。在2.2节中已详细介绍了使用四级多尺度池化的MCE模块的结构。而另一种关于MCE模块的设计方案则使用三级多尺度池化,池化后特征的空间维度分别为2×2, 4×4和6×6,剩余部分的设计则与2.2节中介绍的四级多尺度池化MCE模块的结构相同。表3展示了不同池化等级对分类精度的影响,其中MFCE(2, 4, 6, 8)表示采用四级多尺度池化的网络而MFCE(2, 4, 6)表示采用三级多尺度池化的网络。结果表明,MFCE(2, 4, 6, 8)拥有比MFCE(2, 4, 6)更高的分类精度。这一现象表明,对于高分辨率遥感图像这类具有复杂场景布局的图像,更加精细化、更小粒度的

特征对于最终的分类结果有更大的贡献。因此，在本文的网络结构设计中，我们采用了四级多尺度池化结构的MCE模块。

表3 AID数据集和NWPU-RESISC45数据集上采用不同多尺度池化等级的MFCE网络分类结果

Table 3 Results of MFCE network adopting different levels of multiscale pooling on AID dataset and NWPU-RESISC45 dataset

Methods	OA	
	AID (Tr=20%)	NWPU-RESISC45 (Tr=10%)
MFCE (2, 4, 6)	95.16±0.20	92.17±0.28
MFCE (2, 4, 6, 8)	**95.51±0.09**	**92.42±0.20**

3.4.2 上下文编码对网络分类效果的影响

本节考察MCE模块中上下文编码部分对于分类结果的影响，表4展示了不包含上下文编码部分的MFCE网络和原有MFCE网络以及基线方法在不同数据集上的分类结果。可以看到，移除上下文编码部分后，相较于原有的MFCE网络，不包含上下文编码的MFCE在不同数据集上的分类精度出现了较明显的下降。相比于基线方法(Fine-tuned DenseNet-121)，不包含上下文编码的MFCE方法在AID数据集上的分类精度有一定程度的提高，但在NWPU-RESISC45数据集中，分类精度较基线方法略有降低。综合以上可知，仅使用多尺度信息对提高遥感图像场景分类精度的作用有限，需要进一步对不同尺度、不同子区域的空间关系进行编码。上述实验结果充分证明了本文所设计的MCE模块结构的有效性和合理性。

3.4.3 空间上下文信息对分类精度的作用

本节将MCE模块与不同的网络结构融合，用以验证空间上下文信息对遥感图像场景分类任务的作用。我们选取ResNet-18[19]和经典的VGGNet-16网络与MCE模块进行融合。由于ResNet有不同版本，为保证比较的公平性，选取其参数量与DenseNet-121相当的版本。类似于MFCE网络，针对VGGNet-16和ResNet-18，我们移除其全连接层，然后将MCE模块嵌入到该网络最后一层的卷积层之后，对网络输出特征使用全局平均池化运算，将得到的结果输入一层全连接层以产生最后的分类决策。表5分别展示了各方法在AID数据集(训练集比例为20%)上的分类结果。由以上结果可知，使用MCE模块可以提高不同的主干网络的分类精度。本文所提出的MFCE网络展现出了最优的分类精度，这是由于MFCE采用了网络浅层至深层特征复用的机制，而其余的网络模型则未能充分利用不同层的特征信息。

表4 AID数据集和NWPU-RESISC45数据集不同方法结果比较

Table 4 Comparison of different methods on AID dataset and NWPU-RESISC45 dataset

Method	OA	
	AID (Tr=20%)	NWPU-RESISC45 (Tr=10%)
Fine-tuned DenseNet-121	94.75±0.18	91.56±0.21
MFCE without Context Encoding	94.92±0.19	91.52±0.30
MFCE	**95.51±0.09**	**92.42±0.20**

表5 AID数据集上MCE模块与不同主干网络融合及基线方法的分类结果

Table 5 Results of MCE module combined with different backbones and baselines on AID dataset

Method	OA
Fine-tuned VGGNet-16	90.19±0.38
Fine-tuned ResNet-18	93.10±0.35
Fine-tuned DenseNet-121	94.75±0.18
VGGNet-16+MCE (ours)	**91.57±0.26**
ResNet-18+MCE (ours)	**94.08±0.20**
MFCE (ours)	**95.51±0.09**

3.4.4 参数量和计算复杂度分析

本节对所本文提出的MFCE网络的参数量和乘积累加运算(Multiply-Accumulate Operations，MACs)进行分析。其中，乘积累加运算代表模型的运算复杂度。表6分别对比了不同网络结构的参数量和MACs。结果表明，VGGNet-16网络具有最大的参数量。ResNet-18由于网络结构中没有使用多个全连接层，相较于VGGNet-16实现了参数量和MACs值的下降。而DenseNet-121由于采取了轻量级的网络设计方法[13]，在参数量和MACs上都实现了大幅度降低。MFCE，ResNet-18+MCE以及VGGNet-16+MCE的对比结果表明，MCE模块的引入没有显著增加原有网络的参数量和运算量。对比DenseNet-121，MFCE网络的参数量有一定程度的增加，但仍低于VGGNet-16和ResNet-18。同时，当使用本文设计的MCE模块与VGGNet-16进行结合

时,由于没有使用VGGNet-16中的多个全连接层,网络参数量大大减少。综合以上可知,本文提出的MFCE不仅能够提高网络分类精度,而且具有较小的参数量和较低的计算复杂度。

表6　不同网络结构的参数量和乘加累积运算

Table 6　Parameters and MACs of different networks

Methods	Parameters	MACs
VGGNet-16[5]	138.36M	15.48G
ResNet-18[19]	11.69M	1.82G
DenseNet-121[13]	7.98M	2.87G
VGGNet-16+MCE (ours)	15.52M	20.10G
ResNet-18+MCE (ours)	11.98M	2.42G
MFCE (ours)	10.14M	3.91G

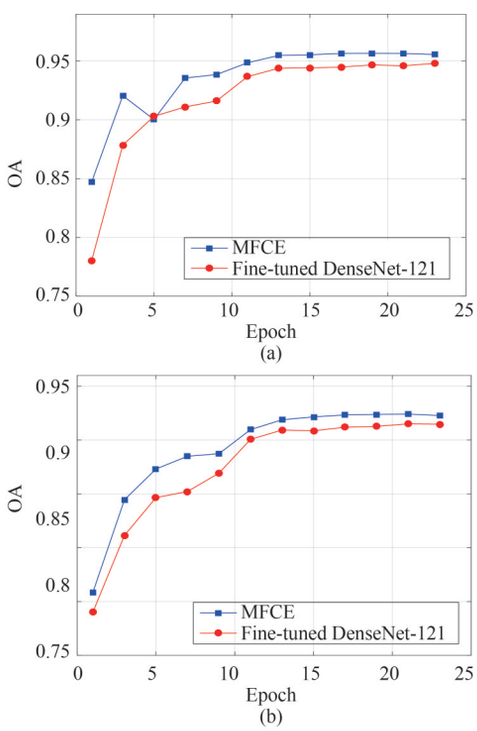

图4　MFCE网络和Fine-tuned DenseNet-121在不同数据集上的测试集精度 (a) AID数据集,(b) NWPU-RESISC45数据集

Fig. 4　Test accuracy with MFCE network and Fine-tuned DenseNet-121 (a) AID dataset, (b) NWPU-RESISC45 dataset

此外,图4展示了MFCE与Fine-tuned DenseNet-121在训练阶段在测试集上分类精度的变化趋势。可以看到,所提出的MFCE网络与基线方法具有相似的收敛速度,同时在分类精度方面与基线方法相比,始终保持着较为显著的提高。

3.4.5　可视化结果

类别激活映射(Class Activation Mappings,CAMs)[20]可以通过热图的形式对神经网络提取到的具有判别性的区域进行高亮显示。为更好理解本文提出的MFCE网络,图5展示了基线方法(Fine-tuned DenseNet-121)和所提出方法的热图输出。图5(a-c)的原始图像来自NWPU-RESISC45数据集,分别属于火车站、立交桥和港口场景类别;图5(d-f)和(g-i)则展示了基线方法和MFCE对应的热图输出。可以看到,在火车站场景中,MFCE网络可以更好地捕捉到站台和铁轨所在的子区域的组合信息,从而实现正确分类;但是,采用基线方法,则没有获取到站台区域的信息,这将导致最终的分类错误。另外,在图5中,还可以观察到MFCE网络普遍拥有更大范围的激活区域,这意味着所提出的网络结构可以更加有效地获取图像不同子区域的信息,进而提高场景级别的分类精度。

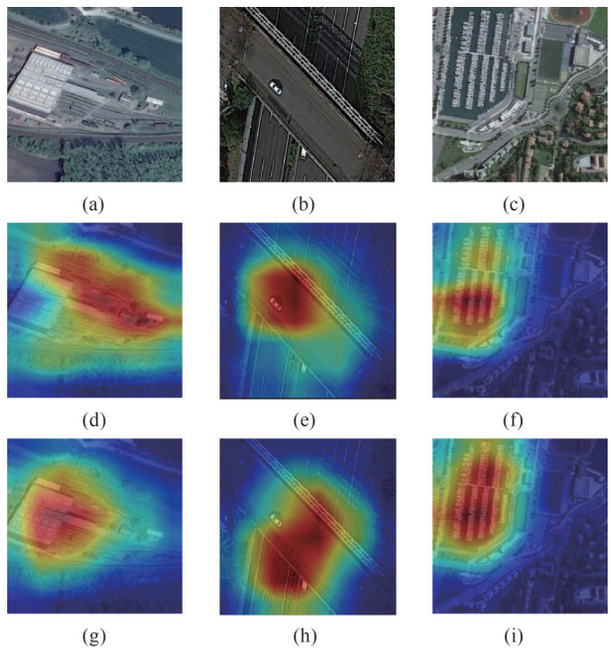

图5　MFCE与Fine-tuned DenseNet-121在NWPU-RESISC45数据集上热图结果对比

注:(d-f)基线方法,(g-i) MFCE网络

Fig. 5　Visual comparison of heatmaps among MFCE and-Fine-tuned DenseNet-121 for NWPU-RESISC45 dataset

Note: (d-f) heatmaps of the baseline, and (g-i) MFCE network

3.5　结果与讨论

由以上实验结果可知,本文所提出的MFCE网络在两个大规模数据集上取得了显著的分类精度提升,并且具有较小的模型参数量和较低的运算复杂度。

在分类精度方面,MFCE网络在AID数据集和NWPU-RESISC45数据集上分别取得了95.51%

(20%的训练样本)和92.42%(10%的训练样本)的分类精度,相较于其他场景分类方法,分类精度分别提升了3.3%和3.2%。在模型参数量方面,本文所提出的网络框架参数量为10.14 M,而遥感图像场景分类任务中最为常用的VGGNet-16网络参数量为138.36 M,远远高于本文方法。在运算复杂度方面,MFCE网络的MACs值为3.91 G,大幅度低于VGGNet-16所需的15.48 G。

4 结论

针对遥感图像场景分类问题,本文提出了一个MFCE网络,在考虑CNN多层特征融合的同时,充分利用图像的空间上下文信息,实现对遥感图像的场景分类。在网络结构设计中,首先,采用具有密集连接机制的主干网络实现浅层到深层的特征融合,从而充分获取图像的空间结构信息和语义信息;然后,该网络采用MCE模块进一步获取遥感图像不同尺度、不同位置的子区域之间的空间关系。通过融合MCE模块和密集连接的主干网络,所提出的MFCE网络可利用CNN中从浅层到深层的多级特征,并充分挖掘蕴含在这些特征中的空间上下文信息。在两个大规模数据集上的实验结果表明,相较于目前的其他遥感图像场景分类方法,本文所提出的网络框架可较好地解决遥感图像场景分类中所存在的图像空间结构复杂这一难点问题,取得了显著的分类精度提升,与此同时,具有较小的参数量和运算复杂度,这些特点将对遥感图像的实际应用具有重要的意义。

References

[1] Hu F, Xia G, Hu J, *et al*. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery [J]. *Remote Sensing*, 2015, **7**(11): 14680-14707

[2] Lowe D. Distinctive image features from scale-invariant keypoints [J]. *International Journal of Computer Vision*, 2004, **60**(2): 91-110

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2005

[4] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks [C]. International Conference on Neural Information Processing Systems, 2012

[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]. International Conference on Learning Representations, 2014

[6] Penatti O A B, Nogueira K, Dos Santos J A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains [C]. IEEE Conference on Computer Vision & Pattern Recognition Workshops, 2015

[7] Cheng G, Yang C, Yao X, *et al*. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs [J]. *IEEE Trans. Geosci. Remote Sens.*, 2018, **56**(5): 2811-2821

[8] Lin T, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2017

[9] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. *Nature*, 2015, **521**: 436-444

[10] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector [C]. European Conference on Computer Vision, 2016.

[11] He N, Fang L, Li S, *et al*. Remote sensing scene classification using multilayer stacked covariance pooling [J]. *IEEE Trans. Geosci. Remote Sens.*, 2018, **56**(12): 6899-6910

[12] Lu X, Sun H, and Zheng X. A feature aggregation convolutional neural network for remote sensing scene classification [J]. *IEEE Trans. Geosci. Remote Sens.*, 2019, **57**(10): 7894 - 7906

[13] Huang G, Liu Z, Van der Maaten L, *et al*. Densely connected convolutional networks [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2017

[14] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]. International Conference on Machine Learning, 2015

[15] Xie J, He N, Fang L, *et al*. Scale-free convolutional neural network for remote sensing scene classification [J]. *IEEE Trans. Geosci. Remote Sens.*, 2019, **57**(9): 6916-6928

[16] Szegedy C, Liu W, Jia Y, *et al*. Going deeper with convolutions [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2015

[17] Xia G, Hu J, Hu F, *et al*. AID: a benchmark data set for performance evaluation of aerial scene classification [J] *IEEE Trans. Geosci. Remote Sens.*, 2017, **55**(7): 3965-3981.

[18] Cheng G, Han J, Lu X. Remote sensing image scene classification: benchmark and state of the art [J]. *Proceedings of the IEEE*, 2017, **105**(10): 1865-1883.

[19] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2016.

[20] Zhou B, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2016.