联合卷积神经网络与集成学习的遥感影像场景分类

余东行，张保明，赵传，郭海涛，卢俊

信息工程大学 地理空间信息学院，郑州 450001

摘 要：针对人工设计的中、低层特征难以实现复杂场景影像的高精度分类以及卷积神经网络依赖大量训练数据等问题，结合迁移学习与集成学习，提出了一种联合卷积神经网络与集成学习的遥感影像场景分类算法。首先基于迁移学习的思想，利用在自然影像数据集上训练好的多个深层卷积神经网络模型作为特征提取器，提取图像多个高度抽象的语义特征；然后构建由Logistic回归和支持向量机组成的Stacking集成模型，对同一图像的多个特征分别训练Logistic模型，将预测概率结果融合构建概率特征；最后利用支持向量机对概率特征训练和预测，得到场景影像的分类结果。利用UCMerced_LandUse和NWPU-RESISC 45两种不同规模的遥感影像数据集进行试验，即使在只有10%的数据作为训练样本情况下，本文方法能够分别达到90.74%和87.21%的分类精度。

关键词：遥感影像，场景分类，卷积神经网络，迁移学习，集成学习

引用格式：余东行，张保明，赵传，郭海涛，卢俊.2020.联合卷积神经网络与集成学习的遥感影像场景分类.遥感学报，24(6)：717-727

Yu D H，Zhang B M，Zhao C，Guo H T and L J. 2020. Scene classification of remote sensing image using ensemble convolutional neural network. Journal of Remote Sensing(Chinese)，24(6)：717-727[DOI：10.11834/jrs.20208273]

1 引 言

遥感影像场景分类是遥感影像解译的重要手段，在灾情监测与评估、目标判读等方面具有重要的应用价值（许凤晖 等，2016）。随着遥感影像分辨率的提高，遥感影像具有更加丰富的空间纹理特征和语义信息，其场景类别呈现多样化、精细化，不同类别的场景影像之间相似性增大，同一类别的场景影像之间差异性也显著增大，导致遥感影像正确分类和识别的难度加大，因此选择更加有效的影像特征表达方式和分类算法是提升场景分类性能的关键。

目前遥感影像场景分类方法主要可分为3类（Cheng等，2017）：无监督图像分类、基于人工设计特征的图像分类和基于深度学习的图像分类。无监督分类方法主要有主成分分析、K-means聚类以及稀疏编码（Sheng等，2012）等，这类方法虽然可以取得较好的分类效果，但不适用于场景的识别。人工设计特征通常可分为低层特征和中层特征，这些特征是利用专业知识所设计的图像局部或全局特征。低层特征如颜色直方图、纹理特征、GIST（Yin等，2015）、梯度直方图特征（Dalal和Triggs，2005）等，这类特征计算简单易于实现，但无法有效表达影像的高层语义信息，分类精度较低。中层特征中最具有代表性的是视觉词袋特征BOVW（Bag of Visual Words），通过聚集、整合低层特征（如SIFT特征），建立起低层特征与高层语义特征之间的联系，来提高图像的分类效果。许多遥感影像场景分类算法也都是在词袋模型的基础上加以改进（Yang等，2010；闫利等，2017），这些基于词袋特征的遥感影像场景分类算法，能够在一定程度上克服利用影像特征点分布生成全局直方图过程中丢失图像局部细节信息的问题，提高词袋特征对遥感影像的信息表达能力，但无法从根本上解决词袋特征存在的单词模糊和冗余问题，难以适应复杂场景影像的高精

收稿日期：2018-07-16；**预印本**：2018-11-07

基金项目：国家自然科学基金(编号：41601507)

第一作者简介：余东行，1993年生，男，硕士研究生，研究方向为摄影测量与遥感、深度学习与遥感影像解译。E-mail：dong_hang@aliyun.com

通信作者简介：张保明，1961年生，男，教授，研究方向为摄影测量与遥感。E-mail：zbm1961@163.com

度分类。基于深度学习的遥感影像场景分类方法目前主要分为两类：(1) 在训练好的卷积神经网络CNN (Convolutional Neural Network) 模型基础上微调 (Fine-tune) (Hu 等，2015；Cheng 等，2017；Zhou 等，2017)，分类精度远高于利用人工设计特征的方法；(2) 将传统人工设计特征与卷积神经网络模型相结合，实现二者优势互补 (何小飞 等，2016；Wang 等，2017；郑卓 等，2018)，这种方法有效提高了单纯利用卷积神经网络模型对遥感影像场景分类的精度。相比传统分类方法，卷积神经网络模型在图像分类上具有巨大的效率和精度优势，是目前图像分类最有效的方法，但其训练依赖于数量巨大的数据、训练过程复杂耗时。标注大量训练数据需要耗费大量的人力物力，这些数据的缺乏是制约其分类性能的主要因素之一。因此研究在较少训练数据情况下，针对相似性场景、复杂场景影像的高精度分类具有较大的理论意义和应用价值。

为了充分利用图像高层语义特征，降低深层卷积神经网络模型对训练数据的依赖程度，提高遥感影像场景分类的效果，本文综合迁移学习与集成学习的优势，提出了一种结合卷积神经网络与集成学习的遥感影像场景分类方法ECNN (Ensemble Convolutional Neural Network)。方法在图像特征的构建方面，利用现有卷积神经网络的预训练模型提取同一图像多个高层语义特征；在分类器的设计方面，采用由Logistic回归和支持向量机组成的Stacking集成分类器，从而实现从特征表达的构建和分类算法的设计两个方面共同提高遥感影像的场景分类精度。

2 基本原理

2.1 卷积神经网络

卷积神经网络采用局部连接和权值共享机制，提取的图像特征具有尺度和平移不变性。卷积神经网络的基本结构一般由卷积层、池化层以及全连接层构成 (周飞燕 等，2017；常亮 等，2016)，如图1所示。

(1) 卷积层：卷积层由若干特征图组成，它是在图像上利用可训练的卷积核进行卷积运算和非线性映射而得到，其计算公式为

$$\boldsymbol{x}_j^l = f\left(\sum_{x_i \in M_{l-1}} \boldsymbol{x}_i^{l-1} * \boldsymbol{k}_{ij}^l + \boldsymbol{b}_j^l\right) \tag{1}$$

式中，$\boldsymbol{x}_j^l$为l层卷积层的第j个特征图，$\boldsymbol{k}_{ij}^l$为l层的卷积核矩阵，M_{l-1}为$l-1$层特征图的集合，$\boldsymbol{b}_j^l$为网络偏置参数，f为激活函数。卷积核用于图像的特征提取，是卷积神经网络模型的主要参数之一，直接影响卷积神经网络模型提取特征的性能。激活函数定义了对数据非线性映射的转换方式，使得卷积神经网络能够更好地解决特征表达能力不足的问题，常用的激活函数有sigmoid、tanh、ReLU等。

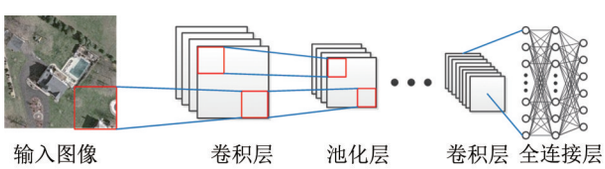


图1 卷积神经网络基本结构

Fig.1 The typical architecture of CNN

(2) 池化层：池化层的操作是对卷积层的图像特征进行降维，并且最大程度保留其显著特征信息，增强特征的平移不变性，缩减下一层特征图的输入大小和参数数量，从而大大降低整个模型的计算复杂度和过拟合的可能。常用的池化操作有最大池化、平均池化、随机池化等。

(3) 全连接层：全连接层的神经元与上一层的节点相连接，整合所有经过多层卷积、池化与非线性操作所得到的特征图，并将2维特征图转化为一维特征向量，该特征向量表示为图像的全局信息，用于图像分类，其计算公式为

$$\boldsymbol{x}^l = f\left(\boldsymbol{w}^l \boldsymbol{x}^{l-1} + \boldsymbol{b}^l\right) \tag{2}$$

式中，$\boldsymbol{w}^l$为全连接层中的权值，$\boldsymbol{b}^l$全连接层l的偏置参数，$\boldsymbol{x}^{l-1}$为前一层的输出特征图。卷积神经网络中普通全连接层的激活函数通常仍采用ReLU等激活函数，但最后一个全连接层为Softmax分类层，用于预测每一类的概率，其表达式为

$$\hat{y} = P(y|\boldsymbol{x}) = \begin{bmatrix} P(y=1|\boldsymbol{x}) \\ P(y=2|\boldsymbol{x}) \\ \vdots \\ P(y=C|\boldsymbol{x}) \end{bmatrix} = \frac{1}{\sum_{i=1}^{C} \mathrm{e}^{\boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{\theta}_i}} \cdot \begin{bmatrix} \mathrm{e}^{\boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{\theta}_1} \\ \mathrm{e}^{\boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{\theta}_2} \\ \vdots \\ \mathrm{e}^{\boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{\theta}_C} \end{bmatrix} \tag{3}$$

根据训练数据$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$，$y_i \in \{1, 2, 3, \cdots, C\}$，建立交叉熵损失函数 (式 (4))，利用梯度下降算法优化损失函数即可求解模型参数θ。

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\left(-\sum_{c=1}^{C}y_i^{(c)}\cdot\log P(y_i=c|\boldsymbol{x}_i,\boldsymbol{\theta})\right)+\lambda R(\boldsymbol{\theta}) \quad (4)$$

式中，$y_i^{(c)}=\begin{cases}1 & ,y_i=c\\0 & ,y_i\neq c\end{cases}$，$R(\boldsymbol{\theta})$为正则化约束条件。

由卷积神经网络的结构可知，网络模型的参数主要集中于卷积层中的卷积核和全连接层之间的连接权重，其中卷积层提取图像的特征，全连接层用于特征的整合和分类。增加卷积神经网络的层数，有助于提高模型的特征提取能力（Simonyan，2014），但当网络加深时，卷积核的参数增多，需要借助于大量数据才能完成训练，在小数据集下训练难以实现参数的全局最优。Krizhevsky等（2012）率先将深层卷积神经网络应用于大规模图像分类问题，精度远远高于采用传统人工设计特征的方法，随后涌现出大量基于卷积神经网络的图像分类算法，广泛应用于计算机视觉领域中图像分类和目标识别等任务。建立一个大规模遥感影像标注数据集不仅耗费巨大的人力物力，同时还面临复杂的训练过程等问题，而解决这些问题的一个有效途径是采用迁移学习策略。

2.2 迁移学习

迁移学习是指利用已有方法和数据去帮助解决具有相似性或相近性任务的方法（Pan和Yang，2010；Weiss等，2016）。深度学习在自然图像下的目标分类和识别方法已渐趋成熟，自然图像下的分类模型应用于遥感影像的特征提取，可在一定程度上解决遥感影像场景分类训练数据缺乏导致难以训练的问题。由卷积层和池化层可以构成一个通用性较强的特征提取器，能够提取图像中高度抽象性的深层特征。一个经过大规模数据集训练好的深层卷积神经网络的卷积层和池化层的参数一旦训练完成，只需要在此基础上对部分参数进行小范围调整，即可适用于其他图像分类任务，这大大减少了整个网络模型的训练参数和所需训练数据。基于迁移学习的图像分类方法一般采用特征迁移和部分参数迁移的方式：保留训练完成的卷积神经网络模型中特征提取层中的所有参数，直接应用于遥感影像的特征提取（Hu等，2015）（图2），将提取到的特征采取其他分类方法进行分类，能够在一定程度上避免训练数据缺乏所导致的过拟合以及训练过程复杂等问题。但遥感影像与常规自然场景影像存在巨大差异，将深层卷积神经网络迁移至遥感影像的场景分类上，必须充分挖掘和完善迁移学习的机制。

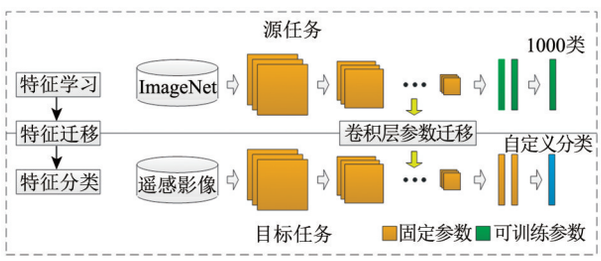


图2 卷积神经网络的迁移学习

Fig.2 Transfer learning using deep CNN

2.3 集成学习

由于不同卷积神经网络模型所提取的特征以及不同分类方法具有明显的差异，单独使用一种特征或者分类方法难以实现高精度分类。集成学习（周志华，2016）通过构建多个分类器以集成方式完成学习任务，不仅能够实现分类器之间的优势互补，从而获得比单一分类器更好的效果，还能减少对训练所需数据的依赖程度。常用的集成学习策略有3种：Bagging、Boosting和Stacking。

Bagging法主要通过在训练阶段对原始训练数据集随机抽样构成子训练集，并用每一个子训练集训练一个分类器，在预测过程中将多个分类器对同一数据的预测结果进行投票或求概率均值，从而得到最终预测结果。Boosting法是一种基于迭代训练的集成方法，在训练过程中对每一个训练子集赋予相同的权重，每次训练后，对训练失败的训练子集赋以较大的权重，使得算法在后续的学习中对比较难的训练子集进行学习，最终预测函数对分类问题采用有权重投票方式。Stacking方法是一种基于多级分类思想的集成学习方法，利用初始训练数据学习出若干子分类器组成第一级分类器（基分类器），并将第一级分类器的预测结果作为新的特征，训练第二级分类器（元分类器）。在理论上，Stacking方法可以用来表示Bagging和Boosting法，同时也具有更优秀的分类性能，因此本文选择Stacking集成方法，其算法如表1所示（Zhou，2012）。

表1 Stacking算法流程

Table 1 The algorithm of Stacking

输入：数据集$\boldsymbol{D}=\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$

由若干分类器组成的第一级分类器：$\boldsymbol{H}=\{H_1, H_2, \cdots, H_m\}$

第二级分类器：L；

训练：

步骤1：利用数据集$\boldsymbol{D}$训练$\boldsymbol{H}$中每一个子分类器；

for t =1 to m do：

Train H_t using $\boldsymbol{D}$

end

步骤2：利用数据集$\boldsymbol{D}$和分类器H生成新的数据集$\boldsymbol{D}_{\mathrm{H}}$：

$\boldsymbol{D}_{\mathrm{H}}=\varnothing$

for i =1 to N do：

for t =1 to m do：

$z_{it}=H_t(\boldsymbol{x}_i)$

end

$\boldsymbol{D}_{\mathrm{H}}=\boldsymbol{D}_{\mathrm{H}}\cup\{(z_{i1}, z_2, \cdots, z_m), y_i\}$

end

步骤3：利用生成的数据集$\boldsymbol{D}_{\mathrm{H}}$训练第二级分类器$L$；

输出：预测结果P, $P=L(\boldsymbol{x}')$, $(\boldsymbol{x}', y)\in\boldsymbol{D}_{\mathrm{H}}$.

3 基于ECNN的遥感影像场景分类

本文提出的ECNN模型主要包括预处理、基于迁移学习与卷积神经网络的遥感影像特征提取、集成学习3个主要步骤，整体流程如图3所示。预处理阶段进行图像归一化和数据扩充；在特征提取步骤中，选择目前自然场景下用于图像分类的几种高精度深层卷积神经网络模型，保留网络模型中除最后分类层之外的所有参数，提取同一图像的不同深层特征；在集成学习阶段，首先构建由基分类器和元分类器所组成的两级分类模型，然后利用基分类器分别对同一图像的不同特征向量进行训练和预测，其预测概率分布结果作为图像新的特征，最后将生成的多个特征集成并利用元分类器训练和预测，得到图像最终分类结果。

3.1 预处理

预处理操作包括数据增强和数据标准化。数据增强用于对有限的数据进行扩充，以提高训练数据的多样性，减小模型泛化误差。常用的数据增强方式主要有旋转、裁剪、添加噪声等，为了验证本文方法在较少训练数据情况下的分类精度，只对图像进行90°、180°和270°共3个方向旋转，将训练数据扩充至原来的4倍。数据标准化用于对图像的大小和灰度标准化操作，目前许多卷积神经网络模型采用图像的大小为224×224像素或299×299像素，因此根据所采用网络模型的输入大小，将所有训练图像和测试图像缩放到需要的尺寸下实验。为了减小图像噪声、灰度变换对特征提取及分类的影响，加快训练收敛速度，将彩色图像$\boldsymbol{I}$各个通道的灰度值归一化到区间［-1，1］，如式（5）所示

$$\boldsymbol{I}=(\boldsymbol{I}-127.5)/127.5 \tag{5}$$

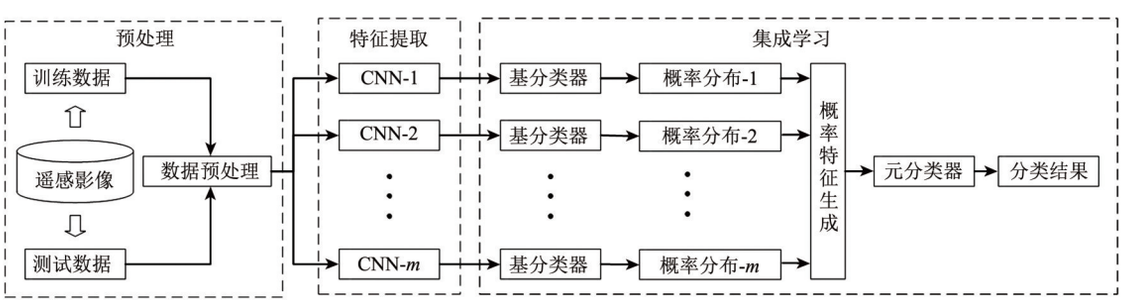


图3 基于ECNN的遥感影像场景分类流程

Fig.3 Flowchart of remote sensing image classification with ECNN

3.2 特征提取与融合

当具有标注信息的遥感影像数据不足以训练一个完整的深层卷积神经网络时，为了更好地迁移学习，选取VGG16（Simonyan和Zisserman，2015）、ResNet50（He等，2016）、Inception（Szegedy等，2015）和Xception（Chollet，2017）4种网络模型进行试验，其模型参数及在ImageNet上的分类精度如表2所示（Chollet，2015）。VGG16模型首次采用更小的卷积核和更深的网络，使其在其他图像数据集上同样具有较好的泛化性能，常常作为图像分类和目标检测的基础模型用来提取特征。VGG16模型包含13个卷积层和3个全连接层，将第二个全连接层输出的4096维向量作为图像提取

的特征。ResNet50在卷积神经网络模型中引入残差机制，在一定程度上解决了当网络模型加深过程中出现梯度消失以及梯度退化问题。ResNet50仅包含一个用于分类的全连接层，移除该全连接层，将输出2048维向量作为提取到的图像特征。Inception和Xception同样只有一个用于分类的全连接层，分别移除Inception、Xception模型中最后的分类层，将提取图像的2048维向量作为该图像的特征。同时可以将4种特征融合，从而获得新的组合特征，图4为利用4种卷积神经网络模型提取特征以及4种特征融合的过程。

表2　4种深层卷积神经网络的模型参数及其在ImageNet数据集的分类精度

Table 2　Brief information of four CNNs on ImageNet

网络模型	输入大小	层数	模型大小/MB	Top-1/%	Top-5/%
VGG16	224×224	23	528	71.5	90.1
ResNet50	224×224	168	99	75.9	92.9
Xception	299×299	126	88	79.0	94.5
Inception	299×299	159	92	78.8	94.4

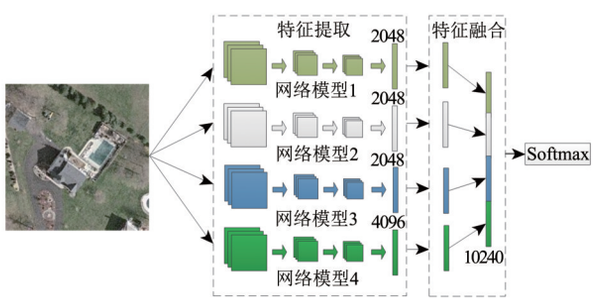


图4　特征提取与融合

Fig.4　Feature extraction and fusion

3.3　集成学习与训练

Stacking本身是一种两级分类器，其模型性能提升的关键在于第一级分类器对训练数据得出差异性较大且预测能力较好的输出值，第二级分类器在第一级分类器的基础上进一步学习，提升预测的准确度和稳定性，而使用复杂分类器进行集成则会产生过拟合的风险。因此Stacking方法倾向于选择更简单的分类方法作为子分类器，以降低过拟合的可能性。Stacking方法中通常采用较为简单的线性Logistic回归等方法来组合模型，本文通过对比几种常用分类方法如Logistic回归、支持向量机SVM（Support Vector Machine）以及Fine-tune在迁移学习所提取特征上的分类精度，最终选择用Logistic回归作为基分类器，SVM作为元分类器构建Stacking模型，其算法实现方式如下。

（1）利用4种网络模型分别对训练图像提取特征，构建迁移特征训练集$\boldsymbol{D}=\{\boldsymbol{D}_1,\boldsymbol{D}_2,\boldsymbol{D}_3,\boldsymbol{D}_4\}$，其中$\boldsymbol{D}_k=\{(\boldsymbol{X}_{k,1},y_{k,1}),(\boldsymbol{X}_{k,2},y_{k,2}),\cdots,(\boldsymbol{X}_{k,N},y_{k,N})\}$，$(\boldsymbol{X}_{k,i},y_{k,i})$表示图像$i$利用卷积神经网络模型$k$所提取特征及对应类别信息，$k\in\{1,2,3,4\}$，$y_{k,i}\in\{1,2,\cdots,M\}$，$M$为图像类别数量，$N$为训练样本总数。

（2）构建由4个Logistic回归模型组成的基分类器$\boldsymbol{H}=\{H_1,H_2,H_3,H_4\}$，利用迁移特征训练集中的4种特征分别训练一个Logistic回归模型。

（3）构建概率特征空间并训练由一个SVM构成的元分类器L。对任意数据$\boldsymbol{X}_{k,i}\in\boldsymbol{D}_k$，用分类器$H_k$预测，得到概率分布$\boldsymbol{P}_{i,k}=[c^1_{i,k},c^2_{i,k},\cdots,c^M_{i,k}]$，其中$c^j_{i,k}$表示样本$i$经过分类器$H_k$所预测其属于类别$j$的概率。然后根据概率分布构建数据集$\boldsymbol{D}_H=\{(\boldsymbol{X}'_1,y_1),(\boldsymbol{X}'_2,y_2),\cdots,(\boldsymbol{X}'_N,y_N)\}$，$\boldsymbol{X}'_i=[\boldsymbol{P}_{i,1},\boldsymbol{P}_{i,2},\boldsymbol{P}_{i,3},\boldsymbol{P}_{i,4}]$。最后用$\boldsymbol{D}_H$训练分类器$L$。

（4）测试过程：对同一图像提取4种特征利用基分类器$\boldsymbol{H}$预测，生成概率特征后，再利用元分类器L预测，最终得到图像类别的预测结果。

4　试验与分析

为验证本文算法的有效性，利用具有代表性的UCMerced_LandUse（Yang和Newsam，2010）和NWPU-RESISC45（Cheng等，2017）两种遥感影像数据集进行试验。试验平台为Intel（R）i7-7700HK处理器、64 G运行内存，并利用NVIDIA GTX1080 10 G显存加速运算，软件采用TensorFlow、Keras搭建卷积神经网络模型。试验均采取K-折交叉验证，$K\in\{1,2,3,\cdots,9\}$，即随机将训练数据中每类样本划分为10等份，任选K份训练，利用剩余的10-K份验证，每次试验进行n次，取分类结果的平均值作为最终实验结果。

4.1　试验数据

UCMerced_LandUse中的数据选自美国地质调查局（USGS）国家城市地图中的航空影像，包含农田、居民区、森林、油罐等21类场景，每类场景由100幅分辨率约为0.3 m、大小为256×256像素的彩色影像组成，共计2100幅，图5为该数据集的部分样本示例。NWPU-RESISC45数据集是目

前公开的一个较大规模遥感影像数据集，涵盖了UCMerced_LandUse数据集中绝大部分类别的同时又增加了岛屿、船只、教堂、发电站等更加详细的场景，将数据类别扩充至45类。每个类别由大小为256×256像素、分辨率为从30 m到0.2 m不等的700张影像组成，共计31500幅，图6为该数据集的部分样本示例。NWPU-RESISC45覆盖100余个国家和地区，包含不同天气、季节、空间分辨率以及遮挡等因素下获取的影像，相比UCMerced_LandUse数据集，场景更加复杂，分类的难度和挑战更大。

由图5、图6可以看出，高分辨率遥感影像的场景类别多样，不同类别的场景影像具有较大的相似性，同一类别的影像具有较大的差异性，如根据居民区的建筑物的稠密程度分为稀疏居民区、中等住宅区、密集住宅区等，同一类别的森林、河流等的影像在颜色和纹理具有较大的差异性，飞机、油罐等场景类别中既包含只有单个目标的图像又包含存在多个目标的图像，以上这些因素增加了其分类的难度。

图5 UCMerced_LandUse数据集部分样本

Fig.5 Samples from each class of UCMerced_LandUse

图6 NWPU-RESISC45数据集部分样本

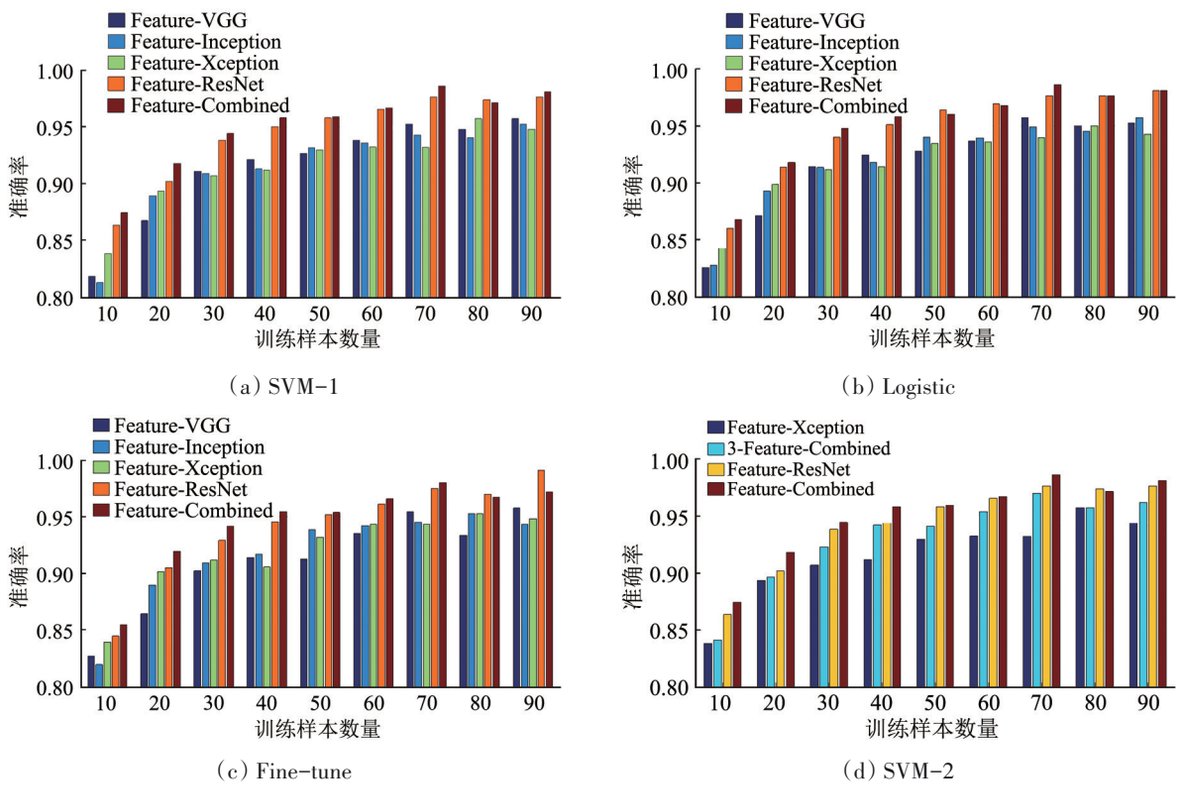
Fig.6 Samples from each class of NWPU-RESISC45

4.2 迁移特征分析

利用模型VGG16、Inception、Xception和ResNet50所提取的特征分别用Feature-VGG、Feature-Inception、Feature-Xception和Feature-ResNet表示，将生成的4种特征融合结果用Feature-Combined表示，对5种特征分别利用SVM、Logistic回归、以及Fine-tune等3种分类方式训练和测试，试验数据为UCMerced_LandUse数据，试验结果如图7所示。

由图7（a）—图7（c）可知，直接利用ImageNet训练好的卷积神经网络模型所提取的特征具有较高的线性可分性，即使在每类场景只有10张训练样本的情况下，不论采取SVM、Logistic回归以及Fine-tune中的哪种方法都可以轻易实现80%以上的分类精度，表明利用卷积神经网络提取的特征具有高度的语义性。由表1可知，4种卷积神经网络模型在ImageNet上的分类精度差别较大，其中精度最好的模型为Xception，而4种卷积神经网络用于遥感影像的特征提取和分类，由ResNet50所提取的特征明显优于其他方法，在较少训练数据情况下（每类样本中10%的数据用于训练，90%的数据用于测试）能够达到85%以上的分类精度，表明自然场景影像与遥感影像具有一定的差异性，ImageNet影像分类的最优模型不一定适用于遥感影像场景分类。在所有特征中，将4种具有差异性的特征进行融合得到的组合特征（Feature-Combined），不论采取哪种分类方法均优

于其他单一特征，表明相对于单个特征而言，通过多种特征融合的方式能够有效提高分类精度，尤其当训练数据较少时（单类样本数量少于50张），提升效果较为明显；随着训练数据增多，样本多样性得到提高，单一特征即可实现较高的分类精度，分类精度趋于稳定，简单特征融合的方式对分类精度难以有较大的提升，这是由于深层卷积神经网络模型所提取的特征具有较高的语义性，当训练样本的多样性满足一定条件后，单纯增加训练数据和特征的维数难以有效提高分类精度，反而可能引入冗余信息或噪声，因此需要从特征设计和分类方法上改善。



（a）SVM-1　（b）Logistic

（c）Fine-tune　（d）SVM-2

图7　不同特征下的分类精度

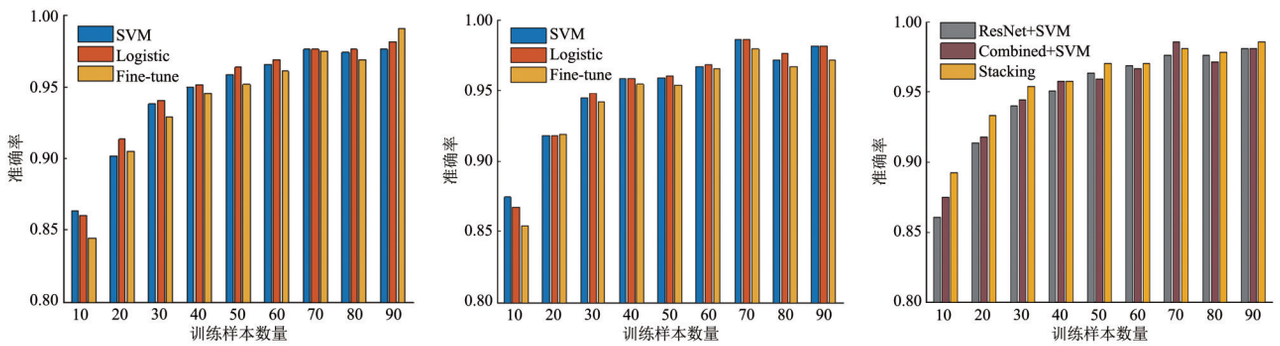
Fig 7　Overall accuracies of different deep features

为了进一步研究特征融合对影像分类精度的影响，将Feature-VGG、Feature-Inception和Feature-Xception等3种精度较差的特征进行融合，得到新的组合特征用3-Feature-Combined表示，采取SVM分类，并与单一特征（Feature-Xception、Feature-ResNet）、组合特征（Feature-Combined）进行比较，试验结果如图7（d）。实验发现，当训练数据较少时（训练数据比例为10%和20%），3种特征融合的方式其分类精度相对于单一特征改善不明显，并且弱于Feature-ResNet特征，表明ResNet50模型所提取的特征优于其他特征，并且在组合的4种特征中，占据主导地位，此外对神经网络提取的特征进行简单的融合，可能存在特征冗余问题。

4.3　集成学习分类精度分析

选取特征组合（Feature-Combined）以及单一特征（Feature-ResNet），对比其在SVM、Logistic回归、Fine-tune以及Stacking集成学习法等4种方法的分类效果（图8）。由图8（a）、图8（b）可知，在少量训练数据情况下（每类样本中训练数据比例为10%和20%），对同一特征而言，采取简单的Logistic回归和SVM的方法甚至能够取得比在原有神经网络模型的基础之上Fine-tune更好的效果，同时Logistic回归和SVM参数更少，训练方法更为简单，优势也更为明显。因此本文Logistic回归模型作为第一级分类器，利用SVM作为第二级分类器建立Stacking集成分类器，对比单一特征和特征融合的分类精度，试验结果如图8（c）所示。当训练数据较少时，Stacking方法比单纯特征融合（Feature-Combined）的分类精度提高约2.5%，当训练数据比例小于50%时，Stacking方法明显优于任何单一特征或组合特征。此外，选择相同的分类

方法作为子分类器，Stacking与Bagging、boosting（本文选用Adaboosting）集成方法的分类精度对比如表3所示，在相同数据集上，Stacking法具有更高的精度。



（a）Feature-ResNet （b）Feature-Combined （c）Stacking

图8不同分类方法结果对比

Fig.8 Overall accuracies of different classifiers

表3 不同集成方法在UCMerced_LandUse的分类精度

Table 3 Accuracies of classification with different ensemble methods on UCMerced_LandUse

/%

集成方法	训练样本所占比例				
	10	20	30	40	50
Stacking	89.26	93.33	95.37	95.79	96.95
Bagging	88.15	92.74	94.56	95.00	96.19
Adaboosting	87.46	91.79	94.42	95.79	95.90

在较少训练数据下，数据扩充能够有效改善影像的分类精度，如表4所示。在不进行数据扩充的情况下，每类样本只有10%用于训练，Stacking方法可以取得的89.26%分类精度；每类样本只有20%用于训练，可以取得93.21%的分类精度，只通过3个方向旋转以扩充数据集，可将精度提高约1.5%。

表4 UCMerced_LandUse数据集下数据扩充对Stacking法分类精度的影响

Table 4 Accuracies of Stacking with data-augment on UCMerced_LandUse

/%

数据扩充	训练样本所占比例				
	10	20	30	40	50
否	89.26	93.33	95.37	95.79	96.95
是	90.74	94.76	96.80	96.92	97.12

为验证本文方法的泛化性，利用场景更加复杂、种类和数量更加丰富的数据集NWPU-RESISC45进行试验。设置两组试验，第1组试验对每类样本数据随机抽取10%数据用于训练，90%数据用于测试。第2组试验随机抽取20%数据用于训练，80%数据用于测试。每组试验进行5次，以其均值作为最终试验结果（表5）。由表5可以得出与数据集UCMerced_LandUse相同的结论：对VGG16和ResNet50等深层卷积神经网络模型进行微调能够取得远高于词袋模型等传统方法的分类精度，其中ResNet50模型特征提取能力最好；多种特征融合的分类精度高于单一特征的分类精度；集成学习方法的分类精度高于特征融合的方式。采取集成学习策略，在不采取数据扩充的情况下，采用10%数据训练，分类精度达到85.75%，采用20%数据训练，分类精度达到88.38%。当仅采用3个方向旋转扩充样本的方式可以将精度提高约1.5%，分别达到87.21%和89.93%，表明本文方法通用性强、可用于少量训练样本情况下遥感影像的场景分类问题。

表5 NWPU-RESISC45数据集的分类精度

Table 5 Overall accuracies on the NWPU-RESISC45

/%

方法	训练样本比例	
	10	20
Color histograms	24.83	25.18
BOVW	33.25	35.41
Fine-tune VGG16	74.97	78.91
Fine-tune ResNet50	83.64	86.43
Feature-Combined(SVM)	84.12	87.46
Stacking	85.75	88.38
Stacking (data augment)	87.21	89.93

4.4 遥感影像分类算法对比

采用MS-DCNN（许凤晖 等，2016）对UCMerced_LandUse数据集样本划分方法，每类样本随机选取80%作为训练数据，其余作为测试数据。为保证试验数据的有效性，将数据随机划分为5份，任选4份作为训练数据，其余一份作为测试数据，进行5组试验，并将5组试验结果的平均值作为最终分类结果（表6和图9）。在相同数据集及划分方法下实验，本文方法总体平均精度达到98.00%，高于其他基于词袋模型以及卷积神经网络的分类算法。如图9所示，UCMerced_LandUse数据集中大多数类别的分类准确率本文方法可以达到100%，分类精度最差的类别为TSC（tennis courts），同样可以达到87%。

表6 不同方法整体分类精度对比

Table 6 Overall accuracies on the UCMerced_LandUse with different methods

模型	平均准确率/%
BOVW（Yang和Newsam，2010）	76.80
MS-DCNN（许凤晖 等，2016）	91.34
CSMCNN（何小飞 等，2016）	92.86
HCV（Wu 等，2016）	91.80
Fisher Vectors（Huang 等，2016）	93.00
GoogLeNet（Nogueira 等，2016）	92.80
GBRCNN（Zhang 等，2016）	94.53
PCANet（Wang 等，2017）	88.46
OVVC-BOW（闫利 等，2017）	87.10
JMCNN（郑卓 等，2018）	88.30
ECNN（本文方法）	98.00±0.18

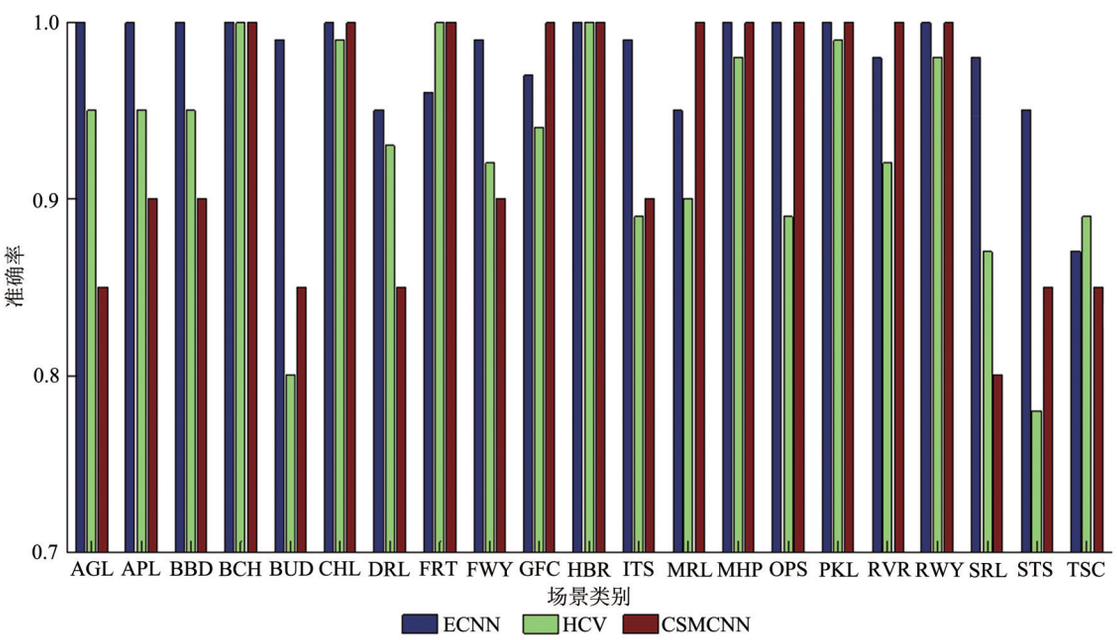


图9 UCMerced_LandUse分类准确率

Fig.9 Classification accuracies of UCMerced_LandUse

5 结 论

本文提出联合卷积神经网络与集成学习的分类方法综合了迁移学习与集成学习的优势，可有效提高当训练数据较少时或深层卷积神经网络难以训练时遥感影像场景分类的精度。相比现有方法，本文方法利用多个预训练卷积神经网络提取影像的特征，增加了特征的多样性和语义性，同时，引入集成学习的方法，从特征的构建和分类器的设计两个方面来提高影像的分类精度。实验结果表明，预训练的卷积神经网络模型所提取的特征具有较高的线性可分性，利用SVM和Logistic回归即可实现不错的分类效果；对训练数据进行扩充、对特征进行融和以及采用集成学习的方式在一定程度上可以进一步提高分类精度。本文方法在遥感影像场景分类的准确性上优于一些仅采用单个卷积神经神经网络的方法。同时，本文也存在一定的不足之处，多个卷积神经网络模型的使用增强了特征的有效性，也增加了特征冗余的可能，并且也不可避免了引入了更多地计算量，

降低了分类的速度。对卷积神经网络模型的合理利用和特征筛选，进一步完善迁移学习机制和集成学习方式，设计出轻量高效、可用于实时处理的遥感影像场景分类方法是需要下一步研究的重点。

参考文献(References)

Chang L, Deng X M, Zhou M Q, Wu Z K, Yuan Y, Yang S and Wang H A. 2016. Convolutional neural networks in image understanding. Acta Automatica Sinica, 42(9): 1300-1312 (常亮, 邓小明, 周明全, 武仲科, 袁野, 杨硕, 王宏安. 2016. 图像理解中的卷积神经网络. 自动化学报, 42(9): 1300-1312) [DOI: 10.16383/j.aas.2016.c150800]

Cheng G, Han J W and Lu X Q. 2017. Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 105(10): 1865-1883 [DOI: 10.1109/JPROC.2017.2675998]

Chollet F. 2015. Keras[EB/OL][2018-05-01]. https://github.com/fchollet/keras

Chollet F. 2017. Xception: deep learning with depthwise separable convolutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 1800-1807 [DOI: 10.1109/CVPR.2017.195]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]

He K M, Zhang X Y, Ren S Q, Sun J and Research M. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He X F, Zou Z R, Tao C and Zhang J X. 2016. Combined saliency with multi-convolutional neural network for high resolution remote sensing scene classification. Acta Geodaetica et Cartographica Sinica, 45(9): 1073-1080 (何小飞, 邹峥嵘, 陶超, 张佳兴. 2016. 联合显著性和多层卷积神经网络的高分影像场景分类. 测绘学报, 45(9): 1073-1080) [DOI: 10.11947/J.AGCS.2016.20150612]

Hu F, Xia G S, Hu J W and Zhang L P. 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing, 7(11): 14680-14707 [DOI: 10.3390/Rs71114680]

Huang L H, Chen C, Li W and Du Q. 2016. Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors. Remote Sensing, 8(6): 483 [DOI: 10.3390/Rs8060483]

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc.: 1907-1105

Nogueira K, Penatti O A B and dos Santos J A. 2016. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recognition, 61: 539-556 [DOI: 10.1016/j.patcog.2016.07.001]

Pan S J and Yang Q. 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10): 1345-1359 [DOI: 10.1109/Tkde.2009.191]

Sheng G F, Yang W, Xu T and Sun H. 2012. High-resolution satellite scene classification using a sparse coding based multiple feature combination. International Journal of Remote Sensing, 33(8): 2395-2412 [DOI: 10.1080/01431161.2011.608740]

Simonyan K and Zisserman A. 2015. Very Deep convolutional networks for large-scale image recognition//Proceedings of 2015 IEEE International Conference on Learning Representations. San Diego, CA, USA: IEEE:1-14

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2015. Rethinking the inception architecture for computer vision//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 2818-2826 [DOI: 10.1109/CVPR.2016.308]

Wang J, Luo C, Huang H Q, Zhao H Z and Wang S WQ. 2017. Transferring pre-trained deep cnns for remote scene classification with general features learned from linear PCA network. Remote Sensing, 9(3): 225 [DOI: 10.3390/Rs9030225]

Weiss K, Khoshgoftaar T M and Wang D D. 2016. A survey of transfer learning. Journal of Big Data, 3: 9 [DOI: 10.1186/S40537-016-0043-6]

Wu H, Liu B Z, Su W C, Zhang W C and Sun J G. 2016. Hierarchical coding vectors for scene level land-use classification. Remote Sensing, 8: 436 [DOI: 10.3390/Rs8050436]

Xu S H, Mu X D, Zhao P and Ma J. 2016. Scene classification of remote sensing image based on multi-scale feature and deep neural network. Acta Geodaetica et Cartographica Sinica, 45(7): 834-840 (许凤晖, 慕晓冬, 赵鹏, 马骥. 2016. 利用多尺度特征与深度网络对遥感影像进行场景分类. 测绘学报, 45(7): 834-840) [DOI: 10.11947/J.AGCS.2016.20150623]

Yan L, Zhu R X, Liu Y and Mu N. 2017. Scene classification of remote sensing images by optimizing visual vocabulary concerning scene label information. Journal of Remote Sensing, 21(2): 280-290 (闫利, 朱睿希, 刘异, 莫楠. 2017. 顾及遥感影像场景类别信息的视觉单词优化分类. 遥感学报, 21(2): 280-290) [DOI: 10.11834/Jrs.201761971]

Yang Y and Newsam S. 2010. Bag-of-visual-words and spatial extensions for land-use classification//Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose, California: ACM: 270-279 [DOI: 10.1145/1869790.1869829]

Yin J H, Li H and Jia X P. 2015. Crater detection based on GIST features. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(1): 23-29 [DOI: 10.1109/Jstars.2014.2375066]

Zhang F, Du B and Zhang L P. 2016. Scene Classification via a gradient boosting random convolutional network framework. IEEE Transactions on Geoscience and Remote Sensing, 54(3): 1793-

1802 [DOI: 10.1109/Tgrs.2015.248868]

Zheng Z, Fang F, Liu Y Y, Gong X, Guo M Q and Luo Z W. 2018. Joint multi-scale convolution neural network for scene classification of high resolution remote sensing imagery. Acta Geodaetica et Cartographica Sinica, 47(5): 620-630 (郑卓, 方芳, 刘袁缘, 龚希, 郭明强, 罗忠文. 2018. 高分辨率遥感影像场景的多尺度神经网络分类法. 测绘学报, 47(5): 620-630) [DOI: 10.11947/J.AGCS.2018.20170191]

Zhou F Y, Jin L P and Dong J. 2017. Review of convolutional neural network. Chinese Journal of Computers, 40(6): 1229-1251 (周飞燕, 金林鹏, 董军. 2017. 卷积神经网络研究综述. 计算机学报, 40(6): 1229-1251) [DOI: 10.11897/SP.J.1016.2017.01229]

Zhou W X, Newsam S, Li C M and Shao Z F. 2017. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. Remote Sensing, 9(5): 489 [DOI: 10.3390/Rs9050489]

Zhou Z H. 2012. Ensemble Methods: Foundations and Algorithms. New York: CRC Press: 83-84

Zhou Z H. 2016. Machine Learning. Beijing: Tsinghua University Press (周志华. 机器学习. 北京: 清华大学出版社, 2016)

Scene classification of remote sensing image using ensemble convolutional neural network

YU Donghang, ZHANG Baoming, ZHAO Chuan, GUO Haitao, LU Jun

Information Engineering University, Zhengzhou 450001, China

Abstract: Scene classification and recognition of remote sensing image is an important task for image interpretation. High-resolution remote sensing images have rich spatial texture features and semantic information, and their scene categories are diversified. As a result, images in the same category have a huge difference and some images in different categories become similar. which makes images difficult to be classified and recognized correctly. Therefore, choosing effective features and classification algorithms can improve classification performance. In this case, high-precision classification can only be achieved by selecting effective features and classifiers.

Traditional scene classification algorithms adopt low-level or mid-level handcrafted features. These features have poor ability to represent high-level semantic information of images, which makes it difficult to achieve satisfactory results on massive complex scene images difficult. Deep learning, especially convolutional neural networks, has made great progress in computer vision. Compared with the methods using handcrafted features, deep learning is currently the most effective way for image classification. The application of a convolutional neural network to remote sensing image classification has achieved higher precision than methods using traditional handcrafted features. However, training a deep convolutional neural network that has too many parameters needs many labeled images, and the process of training is complicated and time-consuming. Generally, a deep convolutional neural network would not perform well with only a few images.

A method for image classification using an ensemble convolutional neural network is proposed to improve the performance of convolutional neural networks. The method is composed of three main phases, namely, preprocessing, feature extraction, and ensemble learning. Firstly, the preprocessing stage includes geometry normalization, image intensity normalization, and image augmentation. Secondly, the feature extraction phase considers several deep convolutional neural networks, which have been well pre-trained on ImageNet, and are chosen to remove the last classification layer in the network and to extract different deep features of the same image. Thirdly, a stacking model is constructed in the ensemble learning phase. The stacking model consists of base and meta classifiers. The base classifier is composed of several logistic regression modes that are used to train different features extracted by deep convolutional neural networks. The meta classifier is a support vector machine. Finally, the probability distribution predicted by the base classifier is used to construct a new dataset that would be trained by the meta classifier.

Experiments were conducted on two datasets named UCMerced_LandUse and NWPU-RESISC45 to verify the effectiveness of the proposed method. Compared with state-of-the-art methods, the proposed method performed better in overall accuracies. The proposed method could greatly improve performance and achieve overall accuracies of 90.74% and 87.21% on the two datasets, respectively, even with only 10% data used for training.

With transfer learning, the features extracted by the deep convolutional neural networks are highly abstract and semantic, which have better ability in classification than other handcrafted features. Through feature fusion and model transferring, the advantages of different features and classification methods could be synthetically utilized. In this way, high classification accuracy could be achieved even with very little training data.

Key words: remote sensing image, scene classification, convolutional neural network, transfer learning, ensemble learning

Supported by National Natural Science Foundation of China (No.41601507)