**Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery**

**摘要：**将CNN框架应用到高分辨率遥感影像场景识别。本文提出两种从CNN不同的层中生成影像特征的方案，第一种，从全连接层中提取激活向量，作为最终的影像特征；第二种，从多尺度对最后一个卷积层提取密集特征，并通过常用的特征编码方法将此特征编码到全局影像特征。

In the first scenario, the activation vectors extracted from fully-connected layers are regarded as the final image features; in the second scenario, we extract dense features from the last convolutional layer at multiple scales and then encode the dense features into global image features through commonly used feature coding approaches.

近几年，高分辨率遥感影像场景分类进展缓慢，主要是因为一般方法无法提供足够的特征表达，而词袋模型BOW与一些无监督特征提取方法只能够建立“中层次”特征表达。而更高级的特征，是较低层次特征的抽象，能有效进行场景分类。

很难通过一个小的样本数据集来训练出高性能的卷积神经网络。

However, it is difficult to train a high-powered deep CNN with small datasets in practice.

很多研究表明，从使用大型数据集预先训练好的CNNs中学习到的中间激励层可以转移到其他有限数据集的识别任务中。

At present, many recent works have demonstrated that the intermediate activations learned with deep CNNs pre-trained on large datasets such as ImageNet can be transferable to many other recognition tasks with limited training data.

**提出问题：**是否可以将与预训练的CNNs成功转移到高分辨率遥感影像的场景分类中，而高分辨率遥感影像场景分类同样具有有限的数据集。

**方案：**

本文通过移除CNNs的最后一层，将剩下的部分作为一个特征提取器。提出两种方案来从不同层提取CNNs的特征：

1. 在整个图像场景中计算CNN的激活向量，并将全连接层中的激活向量视作场景的全局特征表达。

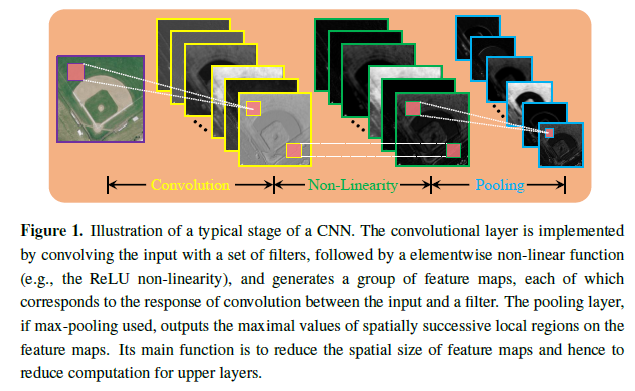
we simply compute the CNN activations over the entire image scene and regard the activation vectors of the fully-connected layer as the global feature representations for scenes;

1. 最后一个卷积层具有原始影像场景的多个尺度，首先从其中生成密集的CNN激活向量，然后使用传统的特征编码方式将密集的卷积特征聚合到一个全局特征表达中，如BOW和Fisher编码。最终，密集的CNN激活向量就可以描述多尺度的空间信息。

we first generate dense CNN activations from the last convolutional layer with multiple scales of the original input image scenes, and then we aggregate the dense convolutional features into a global representation via the conventional feature coding scheme, e.g., the BOW and Fisher encoding. These dense CNN activations describe multi-scale spatial information.

**深度卷积神经网络介绍：**

典型的卷积神经网络有多个级联层次组成。前几个阶段是由卷积层与池化层组成（下图）。卷积层输出特征映射，其每个元素都是通过计算输入特征映射层的某个局部区域与权重（滤子或卷积核）的点积获得的。通常，还要在这些特征映射层上增加非线性的激活函数。池化层进行了一步降采样操作。全连接层出现在多个堆叠的卷积层与池化层之后，最后一个全连接层是一个Softmax层，计算每个类的得分。CNNs通过前向反馈的方式将输入的原始像素值转换成最终的类别得分。而CNNs的参数则通过反向传播算法训练得到。



1. AlexNet：

AlexNet一共由5个卷积层和三个全连接层组成，卷积层中第一、二和五个其后跟随池化层。AlexNet的成功之处在于用到了一些精妙的细节，比如ReLU非线性化、数据扩张及Dropout等。

ReLU是一种半波整流函数能够显著加快训练过程。数据扩张能够通过裁剪、翻转等操作增加样本数量，减少过拟合。Dropout策略即对于每个隐含层，以50%的概率将其输出结果设置为0，这样不再对前向和后向传播过程产生影响，即相当于对不同的网络结构进行了训练，最终求得的参数会适应不同的网络结构，提高了泛化能力， AlexNet在最后两个全连接层使用了这个策略。

1. CaffeNet (Convolutional Architecture for Fast Feature Embedding)

与AlexNet有以下区别：（a）无数据扩张（b）交换了池化层与归一化（归一化是激活函数的作用之一？）层的顺序。

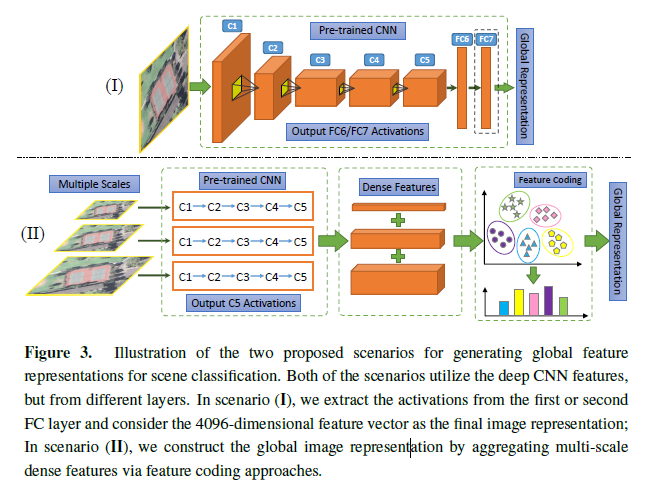
1. VGGNet（基于caffe）
2. VGG-F（快速CNN结构）与AlexNet相近，二者主要区别是VGG-F的卷积核更小，某些卷积层的步长更小。
3. VGG-M（中速CNN结构）其第1个卷积层的步长和池化大小较小，为了平衡运算速度，第4个卷积层的卷积核较小。
4. VGG-S（慢速CNN结构）是OverFeat框架精确模型的简化版本，保留原始OverFeat模型6个层中的前5个卷积层，且第5层的卷积核数量相对较少。相比于VGG-M，其第2个卷积层的步长更小，而第1和第5个卷积层的池化大小更大。
5. VGG-VD网络

这是一个深层卷积神经网络，包含两种：VGG-VD16（包含13个卷积层和3个全连接层）和VGG-VD19（包含16个卷积层和3个全连接层）。

1. PlacesNet

其与CaffeNet的结构完全相同，但其采用Place数据库（一个大尺度scene-centric数据库，包含205种场景类型）来替代ImageNet数据集进行训练。

**本文两种特征提取方案**



方案一：

去除最后一层全连接层，直接从第一个或第二个全连接层计算得到一个4096维的特征向量，将其作为输入图像的全局特征表达。最终使用一个SVM线性分类器进行场景分类。  
 注意问题：（1）训练时，输入图像的大小是固定的（如224x224），因此，当原始影像的大小大于预定义大小时，不可避免地要使其空间分辨率下降；（2）本文采用的数据扩张策略是 “center + corners with horizontal flips” augmentation strategy，分别从影像的中心和4个角选取规定大小的窗口，并进行水平反转，对这10个图像的激活向量进行平均得到影像的全局特征向量。（3）得到的4096维特征向量应该经过ReLU转换，因此最终的特征向量都是非负的。

方案二：

由全连接层获取的全局特征对方向与旋转并不敏感，因此本文将中间卷积层产生的特征视作密集特征，并通过无序特征编码方法进行融合。

方案二移除了所有的全连接层，从最后一个卷积层输出特征映射。由于去除了全连接层，因此输入图像的大小不再固定，可以从任意大小的图像中提取特征。本文选择多个大小的图像输入到网络中，获取多尺度场景信息。然后通过不同的特征编码方法进行融合，方法包括BOW，LLC (Locality Constrained Linear)，IFK (Improved Fisher Kernel)，VLAD (Vector of Locally Aggregated Descriptors)。