doi: 10.6046/zrzyyg.2021271
引用格式: 张鹏强,高奎亮,刘冰,等. 联合空谱信息的高光谱影像深度 Transformer 网络分类[J]. 自然资源遥感,2022,34(3):27 - 32. (Zhang P Q, Gao K L, Liu B, et al. Classification of hyperspectral images based on deep Transformer network combined with spatial - spectral information[J]. Remote Sensing for Natural Resources, 2022, 34(3):27 - 32.)

# 联合空谱信息的高光谱影像深度 Transformer 网络分类

张鹏强, 高奎亮, 刘 冰, 谭 熊
(中国人民解放军战略支援部队信息工程大学,郑州 450001)

**摘要**: 卷积神经网络中的局部卷积运算无法对高光谱影像中的全局语义信息进行充分学习,因此,基于 Transformer 模型设计了一种新颖的深度网络模型,以进一步提高高光谱影像分类精度。首先,利用主成分分析方法对高光谱影像进行降维处理,并选取像素周围邻域数据作为输入样本,以充分利用影像中的空谱联合信息; 然后,利用卷积层将输入样本转换为序列特征向量; 最后,利用构建的深度 Transformer 网络进行分类。Transformer 模型中的多头注意力机制能够充分利用丰富的判别性信息。试验表明,与现有卷积神经网络模型相比,文章方法能够获得更为优异的分类性能。

**关键词**: 高光谱影像分类; Transformer; 深度学习; 自注意力机制
**中图法分类号**: TP 751 **文献标志码**: A **文章编号**: 2097 - 034X(2022)03 - 0027 - 06

## 0 引言

高光谱遥感技术能够同时获取观测区域内丰富的光谱和空间信息,为地物的精细识别和分类提供了可能[1]。高光谱影像分类旨在为影像中的每个像素赋予唯一的类别标识,生成能够反映地物空间分布信息的分类图,为进一步的分析和应用提供数据支撑。支持向量机(support vector machine, SVM)和随机森林等传统分类器能够直接对高光谱影像进行分类。然而,受到高光谱数据高维非线性和"同谱异物、异物同谱"等因素的影响,传统分类器往往无法获得令人满意的分类效果[2]。

深度学习模型通过构建层次化的网络框架,能够逐层提取出具有高判别性和信息性的深层次特征,从而获得更为优异的分类和识别效果。栈式自编码[3]、深度置信网络[4]、循环神经网络[5]等深度模型被率先应用于高光谱影像分类的研究中,在样本充足的条件下取得了较传统方法更为优异的分类效果。卷积神经网络[6-7]利用卷积运算能够直接处理呈网格结构的图像数据,因此更适用于高光谱影像的处理和分析。为了同时利用高光谱影像中空间和光谱信息,二维和三维卷积神经网络被广泛应用于高光谱影像分类。例如,高奎亮等[8]将二维卷积和 NIN 网络结构相结合,设计了一种新颖深度网络模型,有效提高了分类精度; Li 等[9]利用三维卷积构建了适用于高光谱影像分类的深度模型,以充分利用影像中的深度空谱联合信息。除此之外,主动学习[10]、迁移学习[11]和残差学习[12]等先进学习方法也与卷积神经网络相结合,进一步提高了高光谱影像分类的精度和鲁棒性。

卷积神经网络已经成为了高光谱影像处理和分析中的主流深度模型。然而,其仍然存在一定缺陷:卷积运算无法对长距离特征关系进行建模,无法学习全局语义信息。相比之下,Transformer 模型通过将输入图像转换为序列图像块能够更好地利用大范围内的全局语义信息[13]。最近,Transformer 模型已经在图像分割、目标识别等诸多计算机视觉任务中取得了更为优异的表现[14]。受此启发,本文基于 Transformer 模型设计了一种新颖的深度空谱分类网络,以进一步提高高光谱影像分类精度。

## 1 主要算法

### 1.1 Transformer 模型

基本的 Transformer 模型结构如图 1 所示,包含

**收稿日期**: 2021 - 08 - 30; **修订日期**: 2022 - 01 - 16
**基金项目**: 国家自然科学基金项目"基于深度学习的航空序列遥感影像快速三维重建方法研究"(编号: 41801388)资助。
**第一作者**: 张鹏强(1978 - ),男,博士,副教授,主要从事高光谱数据处理、机器学习研究。Email: zpq1978@163.com。
**通信作者**: 高奎亮(1996 - ),男,硕士研究生,主要从事高光谱数据处理、深度学习研究。Email: gokling1219@163.com。

一个自注意力层和一个前馈神经网络。Transformer 模型的输入和输出均为一个特征向量序列,为了更好地考虑输入的特征向量位置信息,在输入第 1 层 Transformer 前,先对特征向量序列进行空间位置编码,然后和特征向量相加作为输入。Transformer 模型的输出同样为特征向量序列,并作为下一层 Transformer 的输入。

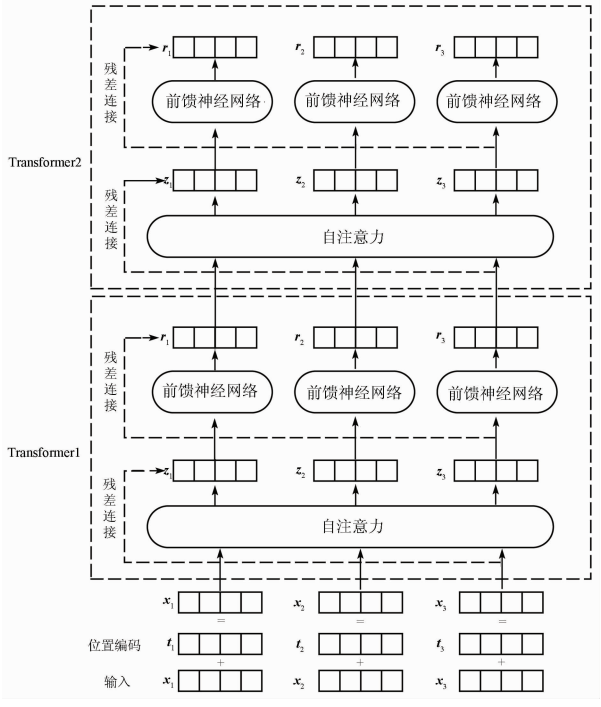



**图 1 Transformer 基本结构**

**Fig. 1 Basic structure of Transformer**

### 1.1.1 空间位置编码

空间位置编码为每个特征向量输出一个维度与特征向量相同的空间位置向量,从而用空间位置向量来描述特征向量的位置关系。本文采用如下的形式对特征向量进行位置编码,即

$$\begin{cases} PE(pos,2i) = \sin\left(\dfrac{pos}{10\ 000^{2i/d_{\text{model}}}}\right) \\ PE(pos,2i+1) = \cos\left[\dfrac{pos}{10\ 000^{(2i+1)/d_{\text{model}}}}\right] \end{cases}, \quad (1)$$

式中: $PE$ 为位置编码; $pos$ 为特征向量在整个序列中的位置; $d_{\text{model}}$ 为特征向量的维度; $i$ 为特征向量的位置。式(1)会在每个特征向量的偶数位置添加 sin 变量,奇数位置添加 cos 变量,以此来产生与原始特征向量维度相同的空间位置向量,然后与原始特征向量相加完成空间位置编码。

### 1.1.2 自注意力层

与卷积神经网络的训练参数卷积核不同,每层 Transformer 的训练参数包含 3 个矩阵 $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V$,这 3 个矩阵分别与输入的向量序列相乘得到查询矩阵、键矩阵和值矩阵。自注意力机制的公式为:

$$\boldsymbol{Z} = Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \quad (2)$$

式中: $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别为查询矩阵、键矩阵和值矩阵; $d_k$ 为输入的维度。为了提高模型的性能,采用多头注意力机制,即使用多个 $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V$ 矩阵生成多个查询矩阵、键矩阵和值矩阵,再根据式(2)输出多个特征值,将多个特征值进行拼接再乘以一个矩阵参数输出最终特征。如图 2 所示, $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$, $\boldsymbol{Z}_3$ 分别是 3 个注意力头输出的特征矩阵(特征序列拼接成特征矩阵),3 个特征矩阵拼接后形成矩阵 $\boldsymbol{Z}$,再与矩阵参数 $\boldsymbol{W}$ 相乘得到最终的输出特征,特征矩阵中每一行为一个特征向量。

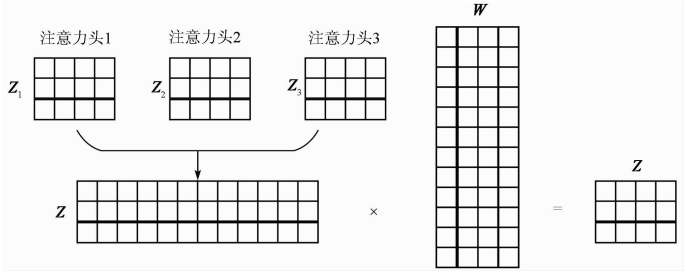



**图 2 多头注意力机制**

**Fig. 2 Multi – head attention mechanism**

为了提高模型的非线性,将自注意力层输出的特征向量序列再分别通过一个前馈神经网络,本文采用 2 层全连接层作为前馈神经网络。另外,在 Transformer 模型中的自注意力层和前馈神经网络层引入残差连接,以提高深度模型的训练效果。

## 1.2 本文网络模型

本文网络模型的整体结构如图 3 所示, Conv2D, TRM 和 MLP 分别代表二维卷积层、Transformer 层和多层感知机。首先,利用主成分分析方法对高光谱数据立方体进行降维处理,并保留前 3 个主成分分量。为了充分利用高光谱影像中的空谱联合信息,选择中心像素周围 32 像素 × 32 像素大小的邻域作为输入样本。具体地,将输入样本沿空间方向划分为 16 个大小相等的图像块; 然后,利用

卷积层将图像块映射为一维特征向量，至此一个输入样本被转换为 16 个一维特征向量；接着，将序列特征向量输入到包含 8 个 Transformer 层的深度网络进行深度特征提取；最后，利用多层感知机输出分类结果。本文网络模型具有端到端的网络结构，以像素领域数据作为输入，以类别标记作为输出。

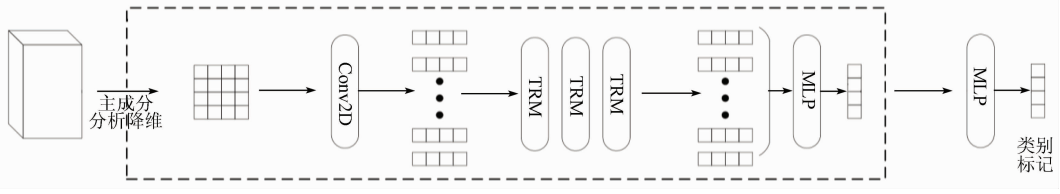



**图 3　本文网络模型**

**Fig. 3　Proposed network model**

## 2　数据集及参数设置

### 2.1　数据集

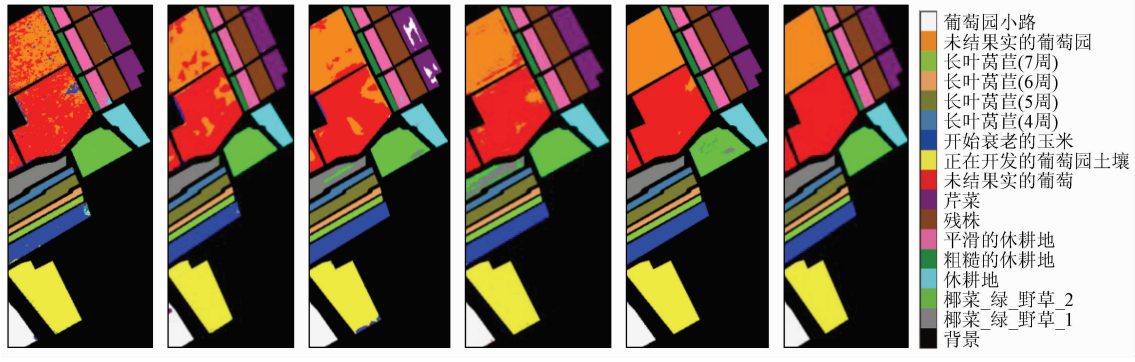
为了验证本文方法的有效性，采用 Salinas 和 Indian Pines 这 2 组高光谱数据集进行试验。Salinas 数据集由 AVIRIS 传感器于 1992 年获取，观测区域为美国加利福尼亚州某山谷，影像大小为 512 像素 × 217 像素，光谱覆盖范围为 0.40 ~ 2.50 μm，空间分辨率为 3.7 m。该数据集共包括野草、休耕地和莴苣等 16 个标注类别和 204 个波段，标注类别集中分布在影像的左侧和上侧，且均为条状和面状地物。Indian Pines 数据集由 AVIRIS 传感器于 2001 年获取，观测区域为美国印第安纳州西北部某处农田，影像大小为 145 像素 ×145 像素，光谱覆盖范围为 0.40 ~ 2.50 μm，空间分辨率为 20 m，共包括 200 个波段可用于分类。该数据集共包括玉米、大豆和树木等 16 个地物类别且在影像中均匀分布，但部分类别包含的标记样本过少。参照相关文献，本文仅选取了 9 个样本数量较多的类别进行实验。另外，2 个数据集均随机选取 200 个标记样本作为训练数据，剩余样本作为测试样本（2 个数据集分别包含 50 929 个和 7 434 个测试样本）。

### 2.2　参数设置

试验中，迭代训练次数设置为 600，学习率设置为 0.000 1，数据批量大小为 64，并利用 Adam 算法进行网络优化，以保证模型进行充分训练。卷积核数量设置为 128，因此每个图像块将被转换为长度为 128 的特征向量。多头注意力机制中头数设置为 8，使模型能够提取到更为丰富的特征。多层感知机中，全连接层的神经元个数分别设置为 128 和 $K$（$K$ 为目标数据集中包含的类别数量）。另外，本文试验的硬件环境为 Intel（R） Xeon（R） Gold 6152 处理器和 Nvidia A100 PCIE 显卡。

## 3　结果与分析

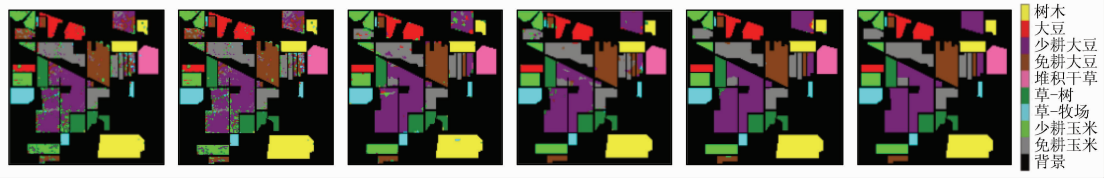
为了验证本文方法的有效性，选择机器学习分类器 SVM、2 种经典的深度学习模型 SSDCNN[15] 和 3D - CNN[9] 以及深度三维残差网络 RES - 3D - CNN[16] 作为对比算法。不同方法在 2 个数据集上的分类结果如图 4—5 所示。可以看到，本文方法的分类结果与真实地面标记最为接近，可以从视觉角度验证了本文方法在高光谱影像分类上的有效性。需要说明的是，图 4（e）中绿色地物区域出现了明显的噪点，但本文方法在其余区域的分类效果明显优




（a） SVM　（b） SSDCNN　（c） 3D - CNN　（d） RES - 3D - CNN　（e） 本文方法　（f） 地面标记

**图 4　Salinas 数据集分类结果**

**Fig. 4　Classification of Salinas dataset**




(a) SVM (b) SSDCNN (c) 3D - CNN (d) RES - 3D - CNN (e) 本文方法 (f) 地面标记

**图 5 Indian Pines 数据集分类结果**

**Fig. 5 Classification of Indian Pines dataset**

于其他方法。本文方法能够通过提高难区分地物的分类精度来改善影像的整体分类效果,从而获得更为平滑的分类结果。

为了进一步对不同方法的分类性能进行定量评价,选择总体分类精度(overall accuray,OA)、平均分类精度(average accuracy,AA)和 Kappa 系数作为评价指标。2 个数据集不同方法的分类结果如表 1—2 所示。从表中可以看到,传统的机器学习分类器 SVM 由于无法利用高光谱影像中的深层次特征,因

**表 1 Salinas 数据集分类结果**

**Tab. 1 Classification results of Salinas dataset**

| | 序号 | 类名称 | SVM | SSDCNN | 3D - CNN | RES - 3D - CNN | 本文方法 |
|---|---|---|---|---|---|---|---|
| 各类别分类精度/% | 1 | 椰菜_绿_野草_1 | 99.20 | 98.90 | 98.75 | 100 | 91.74 |
| | 2 | 椰菜_绿_野草_2 | 99.62 | 89.43 | 99.25 | 98.12 | 100.00 |
| | 3 | 休耕地 | 99.70 | 100 | 99.14 | 99.70 | 100.00 |
| | 4 | 粗糙的休耕地 | 99.50 | 99.93 | 99.50 | 99.86 | 100.00 |
| | 5 | 平滑的休耕地 | 96.75 | 96.56 | 99.74 | 99.81 | 99.81 |
| | 6 | 残株 | 99.42 | 99.97 | 99.62 | 100 | 99.87 |
| | 7 | 芹菜 | 99.36 | 97.18 | 98.21 | 99.55 | 100.00 |
| | 8 | 未结果实的葡萄 | 84.75 | 77.61 | 89.57 | 88.24 | 100.00 |
| | 9 | 正在开发的葡萄园土壤 | 99.10 | 99.11 | 99.98 | 99.50 | 99.98 |
| | 10 | 开始衰老的玉米 | 93.29 | 99.69 | 95.01 | 98.66 | 99.82 |
| | 11 | 长叶莴苣(4 周) | 97.85 | 99.91 | 95.78 | 100 | 100.00 |
| | 12 | 长叶莴苣(5 周) | 99.84 | 100 | 99.79 | 99.95 | 99.95 |
| | 13 | 长叶莴苣(6 周) | 98.58 | 99.89 | 100.00 | 99.45 | 99.89 |
| | 14 | 长叶莴苣(7 周) | 95.79 | 99.53 | 97.99 | 100 | 100.00 |
| | 15 | 未结果实的葡萄园 | 65.74 | 96.68 | 82.12 | 91.98 | 96.40 |
| | 16 | 葡萄园小路 | 98.89 | 99.34 | 99.43 | 99.45 | 100.00 |
| *OA*/% | | | 91.20 | 93.61 | 94.67 | 96.12 | 99.13 |
| *AA*/% | | | 95.46 | 97.11 | 97.12 | 98.39 | 99.22 |
| *Kappa* | | | 0.902 0 | 0.929 1 | 0.940 7 | 0.956 9 | 0.990 3 |

**表 2 Indian Pines 数据集分类结果**

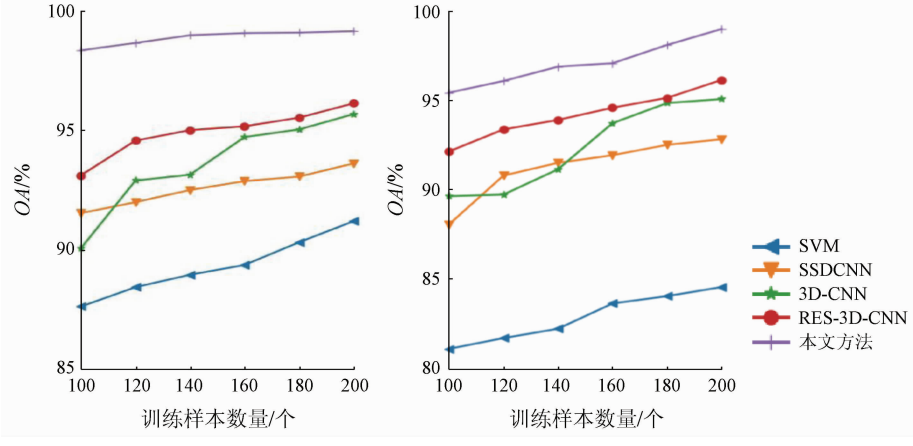
**Tab. 2 Classification results of Indian Pines dataset**

| | 序号 | 类名称 | SVM | SSDCNN | 3D - CNN | RES - 3D - CNN | 本文方法 |
|---|---|---|---|---|---|---|---|
| 各类别分类精度/% | 1 | 免耕玉米 | 75.78 | 88.38 | 93.07 | 94.05 | 99.64 |
| | 2 | 少耕玉米 | 71.12 | 94.34 | 96.27 | 99.04 | 97.42 |
| | 3 | 草 - 牧场 | 89.10 | 97.52 | 98.55 | 98.96 | 100.00 |
| | 4 | 草 - 树 | 96.15 | 99.86 | 97.12 | 99.59 | 97.73 |
| | 5 | 堆积干草 | 99.79 | 100.00 | 100.00 | 100.00 | 100.00 |
| | 6 | 免耕大豆 | 69.98 | 92.49 | 96.81 | 97.02 | 98.37 |
| | 7 | 少耕大豆 | 89.22 | 87.17 | 91.85 | 91.28 | 99.59 |
| | 8 | 大豆 | 79.79 | 99.33 | 97.47 | 99.33 | 97.85 |
| | 9 | 树木 | 99.68 | 96.36 | 98.42 | 99.21 | 98.98 |
| *OA*/% | | | 84.57 | 92.81 | 95.41 | 96.12 | 98.96 |
| *AA*/% | | | 85.62 | 95.05 | 96.62 | 97.61 | 98.84 |
| *Kappa* | | | 0.820 9 | 0.915 9 | 0.946 3 | 0.954 7 | 0.987 8 |

此无法获得令人满意的分类结果。深度学习模型 SSDCNN 和 3D - CNN 分别利用二维卷积和三维卷积进行深度特征提取,分类精度有一定提高。RES - 3D - CNN 利用残差连接和三维卷积核构建了深度残差网络模型,具有更深层次的网络,能够利用深层次的空谱联合特征,因此具有更高的分类精度。在所有对比方法中,本文方法获得了更为优异的分类效果。在 Salinas 数据集上,其 *OA*,*AA* 和 *Kappa* 系数分别较第二名提高了 3.01 百分点,0.83 百分点和 0.033 4;在 Indian Pines 数据集上,其 *OA*,*AA* 和 *Kappa* 系数分别较第二名提高了 2.84 百分点,1.23 百分点和 0.033 1。本文方法通过堆叠 Transformer 模型构建骨干网络,并利用注意力机制和残差连接保证模型能够有效利用到有益于分类任务的深层抽象特征,从而进一步提高分类精度。需要说明的是,本文方法虽然在个别地物上的分类精度略低于 RES - 3D - CNN 模型(例如 Salinas 数据集中第 1 类,Indian Pines 数据集中第 2,4 和 8 类),但其能够明显提高其他方法难以区分的地物的分类精度(例如 Salinas 数据集中第 8 类和第 15 类、Indian Pines 数据集中第 1 类和第 7 类)。这表明,本文方法能够有效提高可分性差的地物的分类精度,从而提升影像的整体分类效果。

另外,深度学习模型需要足够的标记样本进行网络优化和参数更新,然而,实际中获取高质量的标

记样本是十分费时费力的。因此,深度分类模型应对训练样本数量的变化具有良好的适应性。为了探究不同分类方法在训练样本逐渐减少时的分类性能,随机选取每类 100,120,140,160,180,200 个样本作为训练样本进行分类试验,结果如图 6 所示。从图 6 中可以看出,随着训练样本的减少,所有分类方法的分类准确率都逐渐下降。SVM 在 2 个数据集上的分类曲线始终低于其他对比方法。SSDC-NN,3D－CNN 和 RES－3D－CNN 这 3 种基于 CNN 的深度学习分类模型的总体分类精度曲线变化相对平稳,说明它们对训练样本数量的变化具有较好的适应性。本文方法的总体分类精度曲线始终高于其他方法,这表明当训练样本逐渐减少时,该方法具有最好的分类性能。




(a) Salinas 数据集　　(b) Indian Pines 数据集

**图 6　训练样本数量对分类精度的影响**

**Fig. 6　Influence of the number of training samples on classification accuracy**

## 4　结论

为了进一步提高高光谱影像分类精度,基于 Transformer 模型设计了一种新颖的深度分类模型,并展开了相关高光谱影像分类试验,具体结论如下:

1)与基于卷积神经网络的深度学习模型相比,本文方法能够更好地对全局语义信息进行建模,从而获得更高的分类性能。在 Salinas 和 Indian Pines 这 2 个数据集上的试验结果表明,本文方法能够获得比现有卷积神经网络模型更为优异的分类性能。

2)从类一致性的角度看,本文方法的分类结果具有更好的视觉效果,更接近地面真实标记。

3)在逐渐减少训练样本的条件下,本文方法始终能够获得较为优异的分类效果,表明本文方法对训练样本数量具有较好的适应性。

与常规卷积神经网络相比,本文方法利用 Transformer 模型有效提高了高光谱影像分类精度。下一步的工作将结合半监督学习、元学习等方法进一步提高高光谱影像在训练样本受限条件下的分类精度。

**参考文献(References):**

[1] He L,Li J,Liu C,et al. Recent advances on spectral－spatial hyperspectral image classification:An overview and new guidelines [J]. IEEE Transactions on Geoscience and Remote Sensing,2018,56(3):1579－1597.

[2] Ghamisi P,Plaza J,Chen Y,et al. Advanced spectral classifiers for hyperspectral images:A review[J]. IEEE Geoscience and Remote Sensing Magazine,2017,5(1):8－32.

[3] Tao C,Pan H,Li Y,et al. Unsupervised spectral－spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification[J]. IEEE Geoscience and Remote Sensing Letters,2015,12(12):2438－2442.

[4] Li T,Zhang J,Zhang Y. Classification of hyperspectral image based on deep belief networks[C]// Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP),2014.

[5] Zhang X R,Sun Y J,Jiang K,et al. Spatial sequential recurrent neural network for hyperspectral image classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2018,11(11):4141－4155.

[6] Xu Q,Xiao Y,Wang D,et al. CSA－MSO3DCNN:Multiscale octave 3D CNN with channel and spatial attention for hyperspectral image classification[J]. Remote Sensing,2020,12(1):188.

[7] Gao K,Guo W,Yu X,et al. Deep induction network for small samples classification of hyperspectral images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2020,13: 3462－3477.

[8] 高奎亮,张鹏强,余旭初,等. 基于 Network In Network 网络结构的高光谱影像分类方法[J]. 测绘科学技术学报,2019,36(5):500－504,510.
Gao K L,Zhang P Q,Yu X C,et al. Classification method of hyperspectral image based on Network In Network structure[J]. Journal of Geomatics Science and Technology,2019,36(5):500－504,

510.

[9] Li Y, Zhang H, Shen Q. Spectral – spatial classification of hyperspectral imagery with 3D convolutional neural network[J]. Remote Sensing, 2017, 9(1): 67.

[10] Xu X, Li J, Li S. Multiview intensity – based active learning for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(2): 669 – 680.

[11] He X, Chen Y. Transferring CNN ensemble for hyperspectral image classification[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 18(5): 876 – 880.

[12] Mou L, Ghamisi P, Zhu X X. Unsupervised spectral – spatial feature learning via deep residual Conv – Deconv network for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(1): 391 – 406.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Thirty – first Conference on Neural Information Processing Systems, 2017.

[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16[Z]. Transformers for Image Recognition at Scale, 2020.

[15] Yue J, Zhao W, Mao S, et al. Spectral – spatial classification of hyperspectral images using deep convolutional neural networks[J]. Remote Sensing Letters, 2015, 6(4 – 6): 468 – 477.

[16] 刘 冰, 余旭初, 张鹏强, 等. 联合空 – 谱信息的高光谱影像深度三维卷积网络分类[J]. 测绘学报, 2019, 48(1): 53 – 63.
Liu B, Yu X C, Zhang P Q, et al. Deep 3D convolutional network combined with spatial – spectral features for hyperspectral image classification[J]. Acta Geodaetica et Cartographica Sinica, 2019, 48(1): 53 – 63.

# Classification of hyperspectral images based on deep Transformer network combined with spatial – spectral information

ZHANG Pengqiang, GAO Kuiliang, LIU Bing, TAN Xiong

(*Information Engineering University*, *Zhengzhou* 450001, *China*)

**Abstract**: The local convolution operation in convolutional neural networks cannot fully learn the global semantic information in hyperspectral images. Given this, this study designed a novel deep network model based on Transformer in order to further improve the classification precision of hyperspectral images. Firstly, this study reduced the dimensionality of hyperspectral images using the principal component analysis method and selected the neighborhood data around pixels as input samples to fully utilize the spatial – spectral information in the images. Secondly, the convolutional layer was used to transform the input samples into sequential characteristic vectors. Finally, image classification was conducted using the designed deep Transformer network. The multi – head attention mechanism in the Transformer model can make full use of the rich discriminative information. Experimental results show that the method proposed in this study can achieve better classification performance than the existing convolutional neural network model.

**Keywords**: hyperspectral image classification; Transformer; deep learning; self – attention mechanism

(责任编辑: 陈 理)