基于时序自注意力机制的遥感数据时间序列分类

张伟雄[1,2]，唐娉[1]，张正[1]

1. 中国科学院空天信息创新研究院, 北京 100094;
2. 中国科学院大学 电子电气与通信工程学院, 北京 100049

摘 要：遥感影像时间序列为土地覆盖分类研究提供了重要的数据基础，利用深度学习提取时序分类特征一直是研究的热点，而基于循环网络和卷积网络的深度学习模型在训练样本不均衡时往往难以在小样本地类上取得高精度分类结果，针对这一问题，本文引入自然语言处理领域最新的自注意力机制方法用于多光谱遥感时序数据分类。通过对Transformer编码器进行两点改进：（1）在多头注意力前添加特征升维层，提升数据的光谱信息；（2）使用拉伸后降维取代全局最大值池化GMP（Global Maximum Pooling）作为特征维度降维策略。构建基于时序自注意力机制的特征提取网络，与循环网络和卷积网络进行对比，利用公开的多光谱遥感时序数据集评估本文所用方法对于小样本类别精度提高的有效性。实验结果表明本文基于时序自注意力机制构建的特征提取网络能够有效应用于多光谱遥感时序数据分类问题，并对小样本地类分类精度提升有所帮助。

关键词：自注意力机制，深度学习，遥感数据时间序列，土地覆盖分类，不均衡样本

中图分类号：P2

引用格式：张伟雄,唐娉,张正.2023.基于时序自注意力机制的遥感数据时间序列分类.遥感学报,27(8): 1914-1924

Zhang W X, Tang P and Zhang Z. 2023. Time series classification of remote sensing data based on temporal self-attention mechanism. National Remote Sensing Bulletin, 27(8):1914-1924[DOI:10.11834/jrs.20210453]

1 引 言

遥感卫星影像分类是研究土地覆盖与土地利用的基础手段。由于遥感卫星的重访问特性，伴随着海量遥感卫星数据的积累，形成了大量的遥感影像时间序列。遥感影像时间序列能够反映地表地物光谱在一定时间尺度和范围内随时间变化的特性，研究其时序特征建模及分类方法对于有效提高地物分类识别的精度具有重要意义。

由于常规的分类方法，如支持向量机SVM（Support Vector Machines）和随机数森林RF（Random Forest）等传统机器学习方法不能有效提取遥感影像时间序列的时序特征，研究人员开展了一系列以提取时序特征为核心的研究，其中动态时间规整DTW（Dynamic Time Warping）方法有大量研究（Petitjean等，2012；Zhang等，2015；Zhang等，2017；Maus等，2016）。DTW方法通过对两条序列进行特征对齐进而描述序列间的相似性，通过相似性度量进行聚类获得地表分类结果。但DTW是一种对时序特征进行浅层挖掘的方式，其获得有效判别特征的能力有限。

随着深度学习领域的快速发展，一些能够获取深层时序特征的深度学习模型越来越受到青睐，被广泛应用在语音处理、股市房价预测、天气预测等领域，正在逐步引入遥感影像时间序列分类问题中。广泛应用的模型有两类，一类以循环神经网络模型RNN（Recurrent Neural Network）为基础，特别是长短时记忆LSTM（Long Short Term Memory）（Hochreiter和Schmidhuber，1997）网络，循环网络的基本思想是通过刻画一个序列当前时刻与之前时刻的相关性，即前向记忆，来建模时序特征。Ienco等（2017）采用LSTM进行特征学习，在基于像素和基于对象的土地利用分类任务上取得优于传统机器学习方法（SVM，RF）的分类精度；Rußwurm和Körner（2017，2018）使用LSTM在基于像素邻域的植被提取任务中，取得了

收稿日期：2020-10-30；**预印本**：2021-03-30

基金项目：国家自然科学基金(编号:41701399,41971396,41701397)

第一作者简介：张伟雄,研究方向为遥感图像处理。E-mail:zhangweixiong@aircas.ac.cn

通信作者简介：张正,研究方向为遥感图像处理,遥感时间序列分析。E-mail:zhangzheng@aircas.ac.cn

比经典RNN网络和仅用单时相数据更优的效果。另一类深层时序建模的方法是时序卷积网络TempCNN（Temporal Convolutional Neural Network），卷积网络不同于循环网络只能考虑前向时间邻域特征，它沿着时序方向对时间序列数据进行一维时序卷积，具有很强的基于时间邻域特征提取能力。Pelletier等（2019）采用时序卷积网络(TempCNN)，在基于像素的分类任务中，验证了该网络的时序特征提取能力强于不考虑时序—光谱结构的RF、光谱变换、时序变换和循环神经网络等方法；而Zhong等（2019）则在多时相农作物分类任务中验证了基于时序卷积网络优于循环网络和多层感知器MLP（Multi-Layer Perceptron）等方法。深度学习网络虽然能对时间序列数据的时序特征进行深层编码和提取，但深度学习网络对于深层时序特征的学习离不开大量的训练数据，而在遥感影像时间序列分类问题中，由于自然地物的自然生长属性，在一定地理范围内很难保证每个地类都有大量均衡的训练样本。基于循环网络和卷积网络的分类模型虽然在整体分类性能上优于传统的机器学习方法和DTW的浅层特征挖掘，但是仍面临小样本地类精度不高的问题，因此如何提高小样本地类的分类精度一直备受关注。

Vaswani等（2017）提出的基于自注意力机制（Self-Attention Mechanism）的Trasformer网络，该网络不同于循环网络LSTM和卷积网络TempCNN之处在于它是一种能够考虑时间序列全局信息，并自动地关注某些对分类特征提取有重要影响的时序位置进行特征编码的网络（Ganort等，2020），自注意力机制方法的诞生使深度学习领域有了不同于卷积网络和循环网络的第3种基本网络架构，已在自然语言处理领域中取得了统治性的地位。Rußwurm和Körner（2019）首次将自注意力机制引入遥感影像时间序列分类问题中，在类别平衡的实验数据上的结果，表明了自注意力机制的总体分类性能与基于循环网络模型相当的情况下，能够主动地忽视一些云干扰时相，对有云层遮挡的观测具有鲁棒性。基于此，本文通过以Transformer编码器为基础，设计了新的基于时序自注意力机制特征提取器用于多光谱遥感时间序列分类中，并解决小样本地类精度低的问题。本文的实际贡献在于，（1）针对遥感时间序列分类任务中，小样本地类精度不高的问题，引入自然语言处理领域内的自注意力机制与Transformer编码器。（2）针对遥感时间序列数据的特点，对Transformer编码器进行了两点改进：在多头注意力之前添加特征升维层，提升数据光谱信息；使用拉伸后降维替代时序方向的全局最大值池化，构造基于时序自注意力机制的特征提取器。利用自注意力机制能够对不同类别产生不同时序位置的关注，从而增大不同类别间的判别差异，获取样本不均衡时对小样本更强特征表达能力，从而提升小样本精度。（3）通过与目前常用的循环网络和卷积网络方法对比，验证了方法的有效性以及本文对原始网络结构两点改进的必要性。

2 基于时序自注意力机制的遥感数据时间序列特征提取方法

2.1 自注意力机制

注意力机制的本质来自于人类视觉注意力机制。人们视觉在感知东西的时候一般不会是一个场景从到头看到尾每次全部都看，而往往是根据需求观察注意特定的一部分。自注意力机制是自动筛选输入信息中的高价值信息的过程，本质上是一种加权机制，注意力函数可以被描述为一个查询（query）向量和一系列Key-Value向量对到输出向量间的映射，输出向量通过计算Value向量的加权和得到，而与每个Value向量对应的权重是通过计算Query向量和相应的Key向量的相关性得到。

2.2 Transformer编码器

Transformer是自注意力机制实现的一个框架，首次提出是在文本翻译的背景下，使用基于自注意力机制设计的编解码器代替常用的循环神经网络作为Seq2Seq架构（Sutskever等，2014）中的编解码器。Transformer以缩放点积注意力（Scaled Dot-Product Attention）的方式实现自注意力机制的思想，图1左图，其计算具体过程如下：

（1）对输入文本的词嵌入（Mikolov等，2013）向量序列$\boldsymbol{e}\in\mathbb{R}^{T\times d_{model}}$（$T$是序列长度，$d_{model}$是词嵌入向量维度）的每个序列位置$t$的$e^t\in\mathbb{R}^{d_{model}}$，如式（1），应用3个共享参数的转换矩阵（$\boldsymbol{W}^K\in\mathbb{R}^{d_{model}\times d_k}$，$\boldsymbol{W}^Q\in\mathbb{R}^{d_{model}\times d_q}$，$\boldsymbol{W}^V\in\mathbb{R}^{d_{model}\times dv}$）计算得Query-Key-Value三元组（$\mathbf{q}^t$，$\mathbf{k}^t$，$\mathbf{v}^t$），$d_q$，$d_k$，$d_v$是Query-Key-Value向量维度数。在具体实践中通常合并所有的Key，Query，Value向量分别构建为矩阵$\boldsymbol{K}$，矩阵$\boldsymbol{Q}$和矩

阵$\boldsymbol{V}$，如式（2）。

$$\boldsymbol{q}^t = \boldsymbol{e}^t\boldsymbol{W}^Q, \boldsymbol{k}^t = \boldsymbol{e}^t\boldsymbol{W}^K, \boldsymbol{v}^t = \boldsymbol{e}^t\boldsymbol{W}^V \tag{1}$$

$$\boldsymbol{Q} = \boldsymbol{e}\boldsymbol{W}^Q, \boldsymbol{K} = \boldsymbol{e}\boldsymbol{W}^K, \boldsymbol{V} = \boldsymbol{e}\boldsymbol{W}^V \tag{2}$$

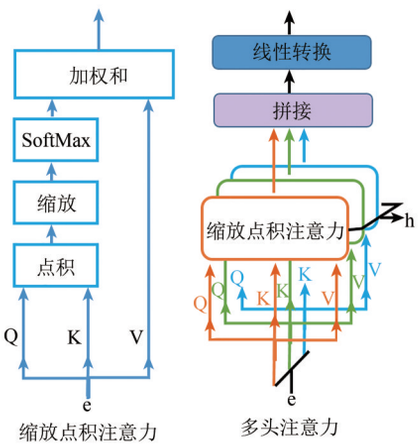


图1 缩放点击注意力和多头注意力

Fig. 1 Scaled dot-product attention and multi-head attention

（2）使用点积作为计算序列间不同序列位置的Key向量和Query向量的相关性的方式，求得不同位置间相互影响的权重矩阵——“注意力”矩阵，并对其进行缩放和SoftMax归一化。

（3）对序列的每个位置进行编码，将（2）中计算得来的权重与Value向量相乘计算加权和，获得序列每个位置的编码输出结果。（2）（3）可用矩阵形式表示为式（3），其中$\sqrt{d_k}$是缩放因子。

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = SoftMax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V} \tag{3}$$

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_1, \cdots, head_n)\boldsymbol{W}^O, head_i = Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \tag{4}$$

一次缩放点积注意力难以关注序列多个重要位置，通过并行地进行多次缩放点积注意力——称之为多头自注意力（Multi-Head Attention），图1右图将h个缩放点积注意力产生的编码输出进行拼接再经过一个线性变换层（$W^O \in \mathbb{R}^{hd_v \times d_{model}}$）变换为一个最终的编码输出，拓展模型关注不同位置的注意力，如式（4）。

Transformer编码器除了多头注意力，还包括位置编码，前馈神经网络，残差连接和归一化等重要操作，如图2展示了一层编码器的结构。

由于缩放点积注意力的计算并没有考虑文本序列的词序信息，因此对词嵌入向量输入进行位置编码，位置编码的要求是每个句子的相同位置的嵌入值是相同的，然后与原来的词嵌入向量相加作为模型新的输入向量。经典的位置编码计算嵌入值按照式（5）—（6）计算：

$$PE(pos, 2i) = \sin(\frac{pos}{N^{2i/d_{model}}}) \tag{5}$$

$$PE(pos, 2i+1) = \cos(\frac{pos}{N^{2i/d_{model}}}) \tag{6}$$

式中，pos为词序位置，i为词向量要素维度，d_{model}为词向量维度数，N是一个控制频率的参数，常取值10000。

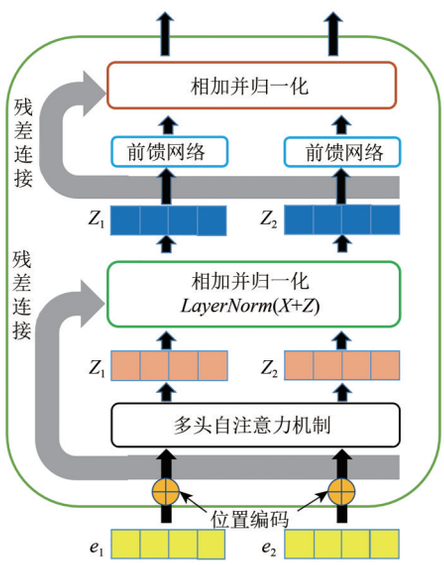


图2 Transformer编码层

Fig. 2 Transformer encoder layer

在多头自注意力后应用逐位置前馈神经网络层，目的是为了巩固网络模型的表达能力。前馈神经网络包含两个线性变换的全连接层，先将输入变换至高维空间再降维至输入维度，中间有一次ReLU激活，将其逐位置地应用到序列上，意味着前馈神经网络在不同位置是参数共享的。如式（7），$z \in \mathbb{R}^{d_{model}}$是经过多头注意力的编码输出，$\boldsymbol{W}_1 \in \mathbb{R}^{d_{model} \times d_{feedforward}}$，$\boldsymbol{W}_2 \in \mathbb{R}^{d_{feedforward} \times d_{model}}$，$d_{feedforward}$是前馈网络中间层神经元数量，且$d_{feedforward} \gg d_{model}$。

$$FFN(z) = \max(0, z\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{7}$$

多个编码层堆叠的深度网络在训练过程中存在梯度消失的问题，使用残差连接和归一化可以有效地消除梯度消失问题。

2.3 基于时序自注意力机制构建遥感数据时间序列特征提取网络

自注意力机制对于序列每个位置来说，通过计算与序列所有位置的相关性来进行编码，考虑了全局的信息，这是其具有很强的特征提取能力的

基础。为了将自注意力机制应用到多光谱遥感数据时间序列分类中，本文以Transformer的编码器为基础，做了适当改造，设计出新的基于时序自注意力机制的特征提取网络，如图3。

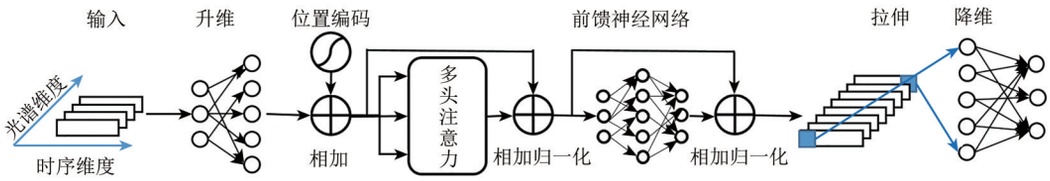


图3 基于时序自注意力机制的特征提取器

Fig. 3 Temporal self-attention mechanism based feature extraction network

Transformer编码器接受的输入是文本经过词嵌入后的的词嵌入张量，有词序维度和特征维度，用于遥感时序数据处理的特征提取器接受基于点像素表示的多光谱遥感时间序列数据张量，有时序维度和光谱维度。由于多光谱遥感数据本身就是多维数据，不同于自然语言处理中的词嵌入，使用一层线性变换进行数据特征升维，目的是为了增强数据的光谱信息，将输入$\boldsymbol{x} \in \mathbb{R}^{T \times d_{\text{input}}}$变换至$\boldsymbol{x} \in \mathbb{R}^{T \times d_{\text{model}}}$。Transformer堆叠了多层编码器，本文只使用了一层编码器。位置编码时仍采用式（5）—（6），则对应的pos是时序位置，i对应光谱维度，d_{model}是升维后的总光谱维度，N取1000。经过多头注意力，位置编码后的输入$\boldsymbol{x} \in \mathbb{R}^{T \times d_{\text{model}}}$编码至$\boldsymbol{z} \in \mathbb{R}^{T \times d_{\text{model}}}$，与输入相加并归一化后，逐位置的经过前馈网络编码至$\mathbf{z} \in \mathbb{R}^{T \times d_{\text{model}}}$，再进行一次残差相加并归一化，在多头注意力和前馈神经网络后使用Dropout。为了与后续的分类器对接并减少高维度带来的“休斯”效应进行特征降维，不同于Rußwurm和Körner（2019）沿着时序维度进行最大值池化特征进行降维，本文将经过Transformer编码器提取的时序特征从二维$\boldsymbol{z} \in \mathbb{R}^{T \times d_{\text{model}}}$拉伸到一维$\boldsymbol{z} \in \mathbb{R}^{Td_{\text{model}}}$并使用一次线性变换降低特征维度$\boldsymbol{z} \in \mathbb{R}^{Td_{\text{model}}} \Rightarrow \boldsymbol{z} \in \mathbb{R}^{d_{\text{model}}}$，作为最终提取的分类判别特征。

整个特征提取器可以调节的超参数包括升维和降维过程中的目标维数d_{model}，多头注意力的头数h，Query-Key-Value向量维度d_q，d_k，d_v，前馈网络中间层维数$d_{\text{feedforward}}$，原则上要保持$d_{\text{feedforward}} \gg d_{\text{model}}$，在本文实验中具体的各超参数取值范围见表1。

3 实验与结果

3.1 实验数据

实验数据来自2017年TiSeLaC时间序列土地覆盖分类竞赛提供的公开数据集。原始数据采集自留尼汪岛2014年全年23景2A级别的Landsat 8影像，研究区具有2866×2633像素大小，30 m空间分辨率和10个波段，包含原始数据前7个波段（Landsat8的Band1到Band7）和3个指数波段（归一化植被指数NDVI，归一化水指数NDWI和亮度指数BI）。对数据逐像素逐波段地通过时序线性插值来替换有云数据。随机采样共得99687个像素构建数据集，分为81714个像素的训练集和17973个像素的测试集，如图4是采样后的像素分布图。参考2012年科林土地覆盖地图（Corine Land Cover）和2014年当地农民填报的土地地块登记结果，将研究区的土地覆盖划分为9个地类。

表1 不同模型的候选超参数

Table 1 Hyperparameter candidates of different models

特征提取模型	可调超参数
LSTM	LSTM层数$L \in \{1, 2, 3\}$
	隐藏状态维数$d_{\text{hidden}} \in \{16, 32, 64, 128, 256\}$
	Dropout概率为$p \in [0, 0.5]$
TempCNN	卷积核大小$k \in \{3, 5, 7\}$
	（相应边缘填充为$padding \in \{1, 2, 3\}$）
	卷积核个数$N \in \{16, 32, 64\}$
	降维层的目标维数$d_{\text{output}} \in \{32, 64\}$
	Dropout概率为$p \in [0, 0.5]$
Transformer	升维和降维的维数$d_{\text{model}} \in \{32, 64\}$
	多头注意力的头数$h \in \{1, 2, 4, 8\}$
	Query-Key-Value向量维$d_q, d_k, d_v \in \{16, 32, 64\}$
	前馈网络中间层维数$d_{\text{feedforward}} \in \{64, 128, 256\}$
	Dropout概率为$p \in [0, 0.5]$

表2是训练集和测试集不同类别的像素数量统计情况，其中Urban Areas，Forests，Sparse vegetation和Rocks and bare soil等4个类别单类占比均超过15%，共计占有样本数75.16%，属于多样本类别，Other built-up surfaces，Grassland，Sugarcane crops，

Other crops 和 Water 等 5 个类别单类占比均不足 10%，是本数据集中的小样本类别，尤其是 Other built-up surfaces，Other crops 和 Water 等 3 个类别，单类占比不足 4%，可见该数据集样本十分不均衡。

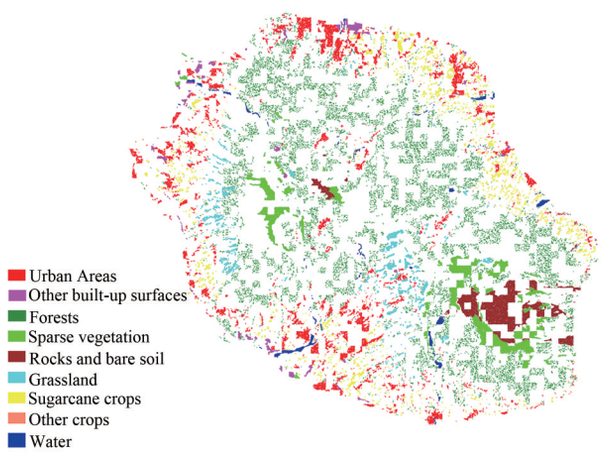


图 4 留尼汪岛 TiSeLaC 数据集像素分布

Fig. 4 Pixels distribution of Reunion Island TiSeLaC dataset

表 2 TiSeLaC 数据集各类别样本统计

Table 2 Per-class Sample statistics in TiSeLaC dataset

类别	训练集	测试集	样本占比/%
Urban areas	16000	4000	20.06
Other built-up surfaces	3236	647	3.89
Forests	16000	4000	20.06
Sparse vegetation	16000	3398	19.46
Rocks and bare soil	12942	2588	15.58
Grassland	5681	1136	6.84
Sugarcane crops	7656	1531	9.22
Other crops	1600	154	1.76
Water	2599	519	3.13
总计	81714	17973	100.0

3.2 对比方法

为了研究自注意力机制在样本不均衡时提高小样本类别分类精度的能力，本文以第 2 节中构建的新的基于时序自注意力机制的特征提取器，在不均衡的样本数据集上，通过与基于循环网络 LSTM 和基于卷积网络 TempCNN 的特征提取器进行分类精度比较，从而验证自注意力机制在全局尺度进行深层时序特征提取对提升小样本类别分类精度的有效性，并通过消融实验验证本文对 Transformer 的相应改进是有效的。

比较验证采用如图 5 的分类网络结构，包括多光谱遥感时间序列数据输入，特征提取器，分类器和类别概率输出，其中特征提取器使用不同特征提取网络，分类器使用多层感知器（MLP）。

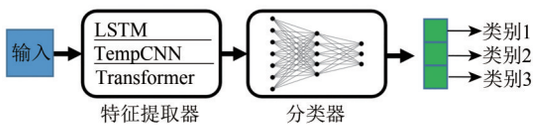


图 5 分类网络结构

Fig. 5 The classification network architecture

3.2.1 循环网络 LSTM

LSTM 在克服 RNN 处理长序列会出现的梯度消失和梯度爆炸的问题的同时，能够沿着时序方向传递每个时间步的状态，获取一定时间范围内的前向时序上下文特征——称为记忆。

图 6 展示了一层 LSTM 网络结构，序列数据 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$，…，$\boldsymbol{x}_T$，按顺序进入 LSTM 单元（Cell）进行处理，LSTM 单元包括 3 个重要的门，输入门决定当前时刻有多少信息被保留，遗忘门决定前面时刻有多少记忆信息被保留，输出门决定当前时刻的记忆有多少被输出到下一时刻。每次 LSTM 单元处理得到当前位置的编码输出和状态，并将状态传递给下一位置，直到序列结束，在这个过程中 LSTM 单元可以循环使用，所以被称为循环网络。这样使得 LSTM 单元能够获取当前时刻前一定距离内时刻的记忆，并且记忆状态和隐藏状态在每个时刻得到更新。

图 7 是本文采用的基于循环网络 LSTM 的特征提取器，该网络通过若干层 LSTM 将遥感时序数据编码到更高维度，不同的 LSTM 层数对特征提取有着不同深度层次的表征。第一层 LSTM 使用原始数据作为输入，堆叠了多层 LSTM 时，之后的下一层 LSTM 使用上一层的隐藏状态作为输入，并在 LSTM 层之间使用 Dropout，各层的首个时刻的隐藏状态都是初始化为零向量，最终将最后一层 LSTM 的最后时刻的隐藏状态作为整个网络特征提取的结果进入分类器。可以调节的网络超参数为 LSTM 层数 L，每层隐藏状态维数 d_{hidden}，具体的各超参数取值范围见表 1。

3.2.2 卷积网络 TempCNN

TempCNN 的核心在于沿着多光谱遥感时序数据的时序维度进行一维卷积提取特征，卷积核的大小决定了感受野的范围，在特征提取时能感知固定的时序邻域范围信息进行编码。

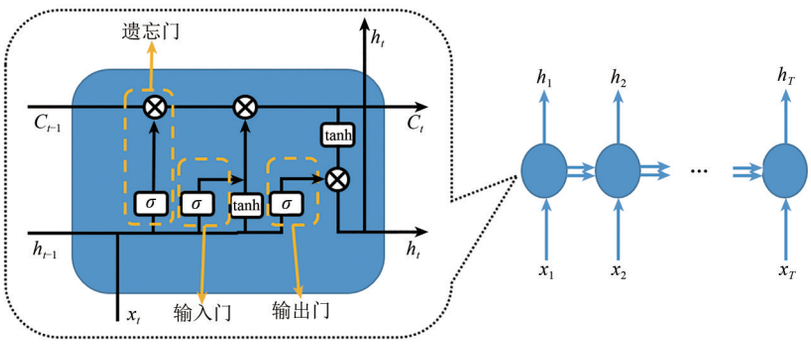


图6 长短时记忆网络

Fig. 6 Long Short Term Memory network

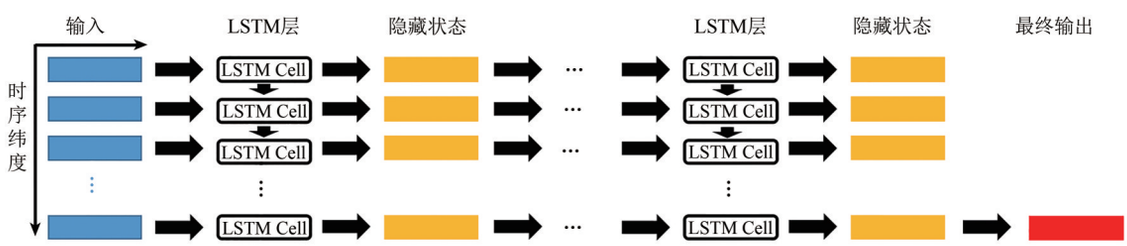


图7 基于LSTM的特征提取网络

Fig. 7 LSTM based feature extraction network

图8是本文采用的基于卷积网络TempCNN的特征提取器，基本沿用了原文的网络结构。先通过堆叠三层时序卷积层进行基于时间邻域特征编码，将遥感时序数据编码至具有时序维度和高特征维度的特征编码结果，每一个卷积层都是通过一维时序卷积，批归一化，ReLU激活和Dropout实现的，一维卷积的步长设为1，并且设置于卷积核相应的边缘填充保持时序长度的完整性。然后通过将编码后的特征结果从二维拉伸为一维向量，再通过一次线性变换降低其维度作为最终的分类判别特征进入分类器。整个模型可以调节的超参数包括卷积核大小k，相应的边缘填充为$padding = \mathrm{int}(\frac{k}{2})$，每层卷积核个数$N$，降维层的目标维数$d_{\mathrm{output}}$，具体的各超参数取值范围见表1。

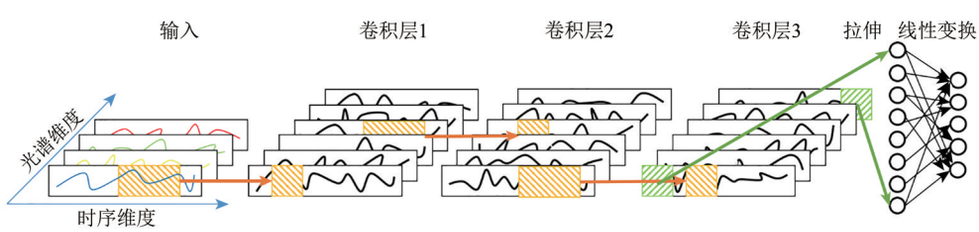


图8 基于TempCNN的特征提取网络

Fig. 8 TempCNN based feature extraction network

3.2.3 时序自注意力机制特征提取网络

在2.3节中，介绍了由Transformer改进的时序自注意力机制特征提取网络，为了验证本文的对Transformer的改进是否有效，增加了对Transformer的消融实验对比组。

对于基础的Transformer结构，记作Basic，没有特征升维步骤，后面的特征降维使用的是全局最大值池化；DimUp则在Basic的基础上增加特征升维；Flatten则在Basic基础上增加拉伸后降维取代全局最大值池化。相应的对比实验组见表3。

3.3 实验设置

实验硬件配置有4块Nvidia Titan XP GPU显卡，32核英特尔至强E5系列CPU，128 GB内存和80 TB硬盘。使用Linux操作系统，实验代码使用Python语言在Pytorch框架下编写。

表3 消融实验

Table 3 Ablation experiments

模型	描述
Basic	原始输入+GMP
DimUp	特征升维+GMP
Flatten	原始输入+拉伸降维
Transformer	特征升维+拉伸降维

为了对比验证基于时序自注意力机制的特征提取网络Tranformer，循环网络LSTM和卷积网络TempCNN在TiSeLaC数据集上的表现，对图5所示的分类网络进行训练时，使用相同的4层MLP分类器，其每层神经元个数为［128，64，32，9］，采用FocalLoss（Lin等，2017）作为损失函数，使用Adam算法（Kingma和Ba，2015）优化损失函数，学习率设置为0.001，指数衰减率设置为［0.9，0.999］。

对分类模型在数据集上进行100次迭代训练和测试，迭代次数足以保证使用不同特征提取器的分类模型可以收敛，迭代过程中批量大小设置为32并随机采样。对每次迭代在训练集上训练出的模型，计算模型在测试集上的总体精度OA（Overall Accuracy），平均交并比mIoU（mean Intersection over Union）和各个类别的单类分类精度。总体精度和平均交并比都反映了模型的整体分类性能，总体精度反映模型对于数据集全体样本的分类精度，但是面对数据集类别不均衡时，总体精度高并不能代表所有类别的单类精度高，本文主要关注小样本类别精度问题，因此使用平均交并比作为反映类别间精度平衡后的整体评价指标。在一次实验中，选取100次迭代中在测试集上mIoU分数最高的迭代作为最后评价模型性能依据。

本文按照表1组合超参数设置，由于深度学习网络模型的学习能力和参数数量有着密切的关联，不同的超参数设置使得模型的参数数量差别也十分巨大，为了合理比较不同模型，本文在表1选择超参数设置的基础上舍弃特征提取网络的参数量超过30万个的超参数组合设置。

3.4 结果与分析

对不同模型在不同超参数下使用随机初始化模型参数，使用100个迭代中mIoU对应的迭代作为一次实验的评价依据，最后对不同模型选取mIoU分数最高的前3个超参数设置实验的结果计算均值作为该模型最终的精度指标结果。

3.4.1 特征提取方法对比结果

表4是基于循环网络LSTM，基于卷积网络TempCNN和本文构建的基于Tranformer编码器的时序自注意力机制特征提取器在相同分类网络架构下，在TiSeLaC数据集上最终的精度评价指标对比，包括各模型在单类上的精度，总体精度OA和平均交并比mIoU。

表4 特征提取方法分类精度表

Table 4 Classification accuracy table of feature extraction methods

类别	LSTM/%	TempCNN/%	Transformer/%
Urban Areas	92.82	**93.13**	92.91
Other built-up surfaces	74.55	75.53	**78.00**
Forests	**92.05**	91.51	91.87
Sparse Vegetation	94.27	**95.60**	95.53
Rocks and bare soil	**97.31**	96.53	96.57
Grassland	89.29	90.64	**91.00**
Sugarcane crops	95.12	94.92	**95.24**
Other crops	67.96	65.15	**70.56**
Water	88.95	88.25	**91.78**
OA	92.56	92.59	**92.98**
mIoU	79.35	79.28	**80.60**

注：最优值加粗表示。

实验结果显示就不同特征提取器分类网络的整体分类性能而言，使用基于时序自注意力机制的特征提取器Transformer优于基于循环网络LSTM和卷积网络TempCNN的特征提取器，总体精度OA达到了92.98%，平均交并比mIoU达到了80.60%，相较于LSTM和TempCNN有1.25%和1.32%的提升。

对于单个类别分类效果而言，Transformer与LSTM，TempCNN在4个多样本类别上表现基本相当，在5个小样本类别上优于LSTM和TempCNN。具体来说，TempCNN在Urban Areas和Sparse Vegetation两个类别上取得最优精度，分别达到93.13%和95.60%，Transformer与之差距甚小，只有0.22%和0.07%。LSTM在Forests和Rocks and bare soil两个类别上取得了最优精度，分别达到92.05%和97.31%，Transformer与之差距也只有0.18%和0.74%。而在其他5个小样本类别上，Transformer均取得了最优精度，在Grassland类别上，Transformer优于LSTM网络1.71%分类精度。尤其是在样本最少的Other built-up surfaces，Other crops和Water等3个类别

上，Transformer取得极其优异的表现，在Other built-up surfaces类别上，分别优于LSTM和TempCNN以3.45%和2.47%，在Other crops类别上优于LSTM和TempCNN以2.6%和5.41%，在Water类别上优于LSTM和TempCNN以2.83%和3.53%。可以看出，深度学习方法对于多样本类别，往往都能够充分挖掘足够的特征信息，因此差异不大，而对于小样本地类精度，时序自注意力机制有着明显的提升。图9是小样本密集分布区域仅显示小样本的分类结果对比图，本文使用的基于时序自注意力机制的方法在小样本类别上的分类效果明显优于循环网络和卷积网络方法。

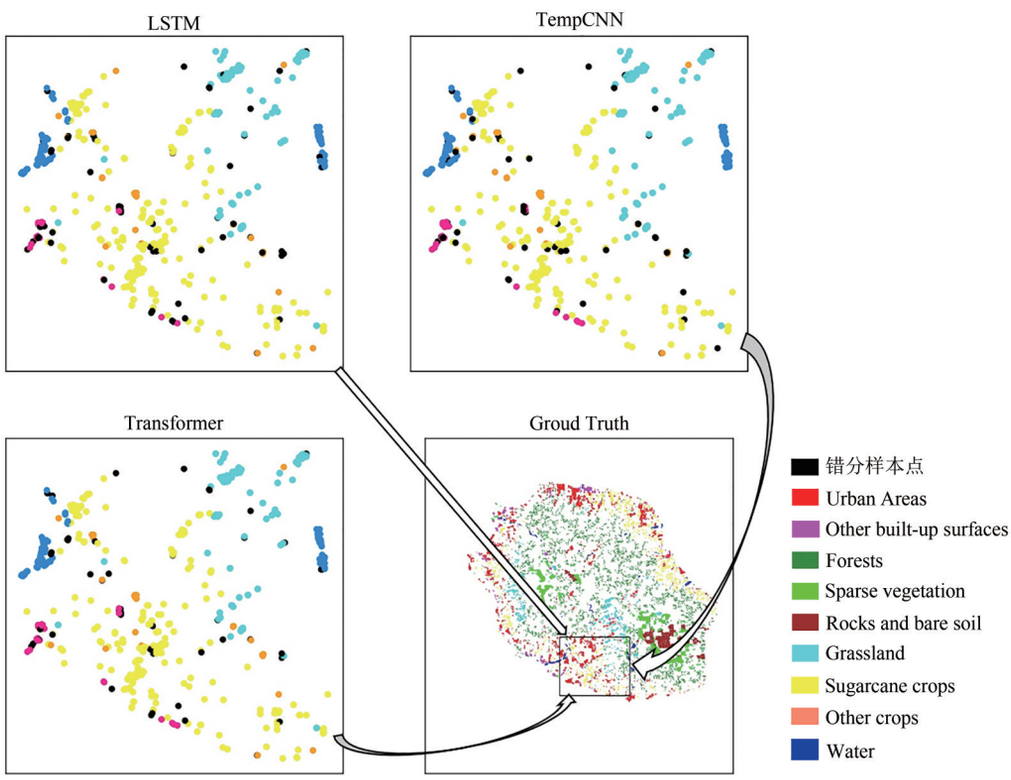


图9 局部区域小样本分类结果

Fig. 9 Small-sample categories classification result in local area

3.4.2 消融实验对比结果

表5是对本文改进后的Transformer编码器进行消融实验的精度对比表，实验结果显示本文对Tranformer编码器所进行的两点改进均是显著有效的。首先，从DimUp与Basic对比可以看出，添加了特征升维层可以提高所有类别的单类精度，尤其是Other crops，基础的Transformer特征提取器几乎不能有效判别该类别，而添加了特征升维之后，使得该类分类精度有了飞跃性的提升，这说明特征升维对于时序自注意力机制处理遥感时间序列数据是十分必要的。同样地，从Flatten与Basic的对比也可以看出，本文所采用的拉伸后降维的降维策略优于全局最大值池化。而同时采用了两种改进的Transformer取得了单类与整体的最优结果。

表5 消融实验分类精度表

Table 5 Classification accuracy table for ablation experiments

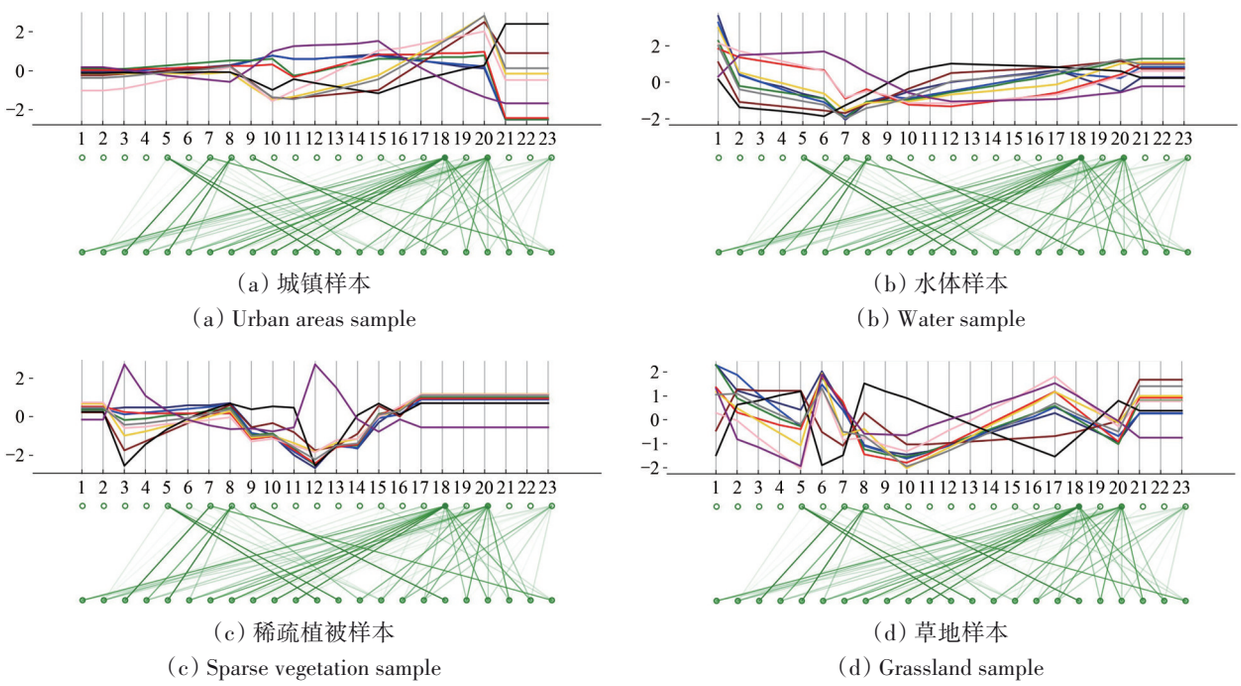
/%

类别	Basic	DimUp	Flatten	Transformer
Urban Areas	89.13	92.42	91.83	**92.91**
Other built-up surfaces	56.36	65.28	65.38	**78.00**
Forests	89.24	90.88	89.49	**91.87**
Sparse Vegetation	86.9	92.90	91.98	**95.53**
Rocks and bare soil	91.79	96.47	94.73	**96.57**
Grassland	73.91	87.93	83.04	**91.00**
Sugarcane crops	88.2	93.16	92.43	**95.24**
Other crops	7.79	58.44	30.95	**70.56**
Water	79.7	87.35	82.15	**91.78**
OA	85.93	91.12	89.50	**92.98**
mIoU	62.92	75.94	70.89	**80.60**

注：最优值加粗表示。

3.4.3 时序自注意力机制可视化

如图10展示了不同类别样本的自注意力权重可视化结果，每个子图上方是各波段归一化时间序列，下方是自注意力机制中时序编码的权重可视化，颜色越深代表权重越大。对于不同类别的样本，时序自注意力机制特征提取器能够对不同时序位置产生不同程度的注意力，这是一种动态的加权机制。相比之下，卷积网络在卷积训练确定之后，对于不同时序位置的编码重要性不存在动态加权；循环网络一定程度上也属于一种加权机制，但是其遗忘门控制前面位置的输入，输入门控制当前位置的输入，随着时序推进前面位置的权重逐渐变小，与当前输入越近，权重越大，因此可以视为一种固定的加权方式。本文所使用的时序自注意力机制能够动态地在不同类别之间关注不同位置，能够有效地扩大类别间的差异性，从而促进小样本地类精度。



（a）城镇样本
（a）Urban areas sample

（b）水体样本
（b）Water sample

（c）稀疏植被样本
（c）Sparse vegetation sample

（d）草地样本
（d）Grassland sample

图10 不同类别样本自注意力权重可视化

Fig. 10 Visual illustration of self-attention weights of samples of different classes

4 结 论

随着遥感技术的发展，海量的遥感影像时间序列数据对对地观测的应用具有重要的意义。本文针对遥感时间序列处理领域目前常用的深度学习分类模型在处理时间序列数据时，需要大量训练数据，而数据中存在样本不均衡的实际情形，使得小样本地类分类精度不高的问题，引入自然语言处理领域最新的方法原理——自注意力机制。基于Transformer编码器进行适当改进构建基于时序自注意力机制特征提取器，借助时序自注意力机制在全局尺度对时序数据进行时序建模，并能自主地关注某些特定时序位置进行特征编码的能力，应用于遥感影像时间序列分类，通过与目前广泛使用的循环网络LSTM和卷积网络TempCNN在公开的遥感时间序列数据集上进行对比。实验结果表明，基于时序自注意力机制方法在全局尺度提取时序特征的方式，在多样本类别上保持了和循环网络使用前向记忆和卷积网络使用时序卷积进行时序特征提取两种方式同水准高精度的同时，能够有效提高小样本类别的精度，明确了基于时序自注意力机制方法用于遥感时间序列数据分类的有效性和意义。

本文将自注意力机制方法用于遥感数据时序

特征建建模，关于自注意力机制方法在遥感时间序列领域的应用，仍有许多值得探索的地方，例如，海量的遥感时序信息提供了丰富的空间–光谱–时序信息，如何进一步基于自注意力机制方法发展混合特征提取建模深度学习网络充满挑战性；在不同的空间分辨率，光谱分辨率和时间分辨率的遥感时序数据中使用自注意力机制应当注意的问题和特点等。

志　谢　本文所用遥感时间序列数据来自Dino Ienco于2017年公开的TiSeLaC竞赛数据集，在此由衷表示地感谢！

参考文献（References）

Garnot V S F and Landrieu L. 2020. Lightweight temporal self-attention for classifying satellite images time series//5th ECML PKDD Workshop on Advanced Analytics and Learning on Temporal Data. Ghent: Springer: 171-181 [DOI: 10.1007/978-3-030-65742-0_12]

Garnot V S F, Landrieu L, Giordano S and Chehata N. 2020. Satellite image time series classification with pixel-set encoders and temporal self-attention//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 12322-12331 [DOI: 10.1109/CVPR42600.2020.01234]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Ienco D, Gaetano R, Dupaquier C and Maurel P. 2017. Land cover classification via multitemporal spatial data by deep recurrent neural networks. IEEE Geoscience and Remote Sensing Letters, 14(10): 1685-1689 [DOI: 10.1109/LGRS.2017.2728698]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//3rd International Conference on Learning Representations. San Diego: ICLR [DOI: 10.48550/arXiv.1412.6980]

Lin T Y, Goyal P, Girshick R, He K M and Dollar P. 2017. Focal loss for dense object detection//2017 IEEE International Conference on Computer Vision. Venice: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Maus V, Câmara G, Cartaxo R, Sanchez A, Ramos F M and de Queiroz G R. 2016. A time-weighted dynamic time warping method for land-use and land-cover mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(8): 3729-3739 [DOI: 10.1109/JSTARS.2016.2517118]

Mikolov T, Sutskever I, Chen K, Corrado G and Dean J. 2013. Distributed representations of words and phrases and their compositionality//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc.: 3111-3119

Pelletier C, Webb G and Petitjean F. 2019. Temporal convolutional neural network for the classification of satellite image time series. Remote Sensing, 11(5): 523 [DOI: 10.3390/rs11050523]

Petitjean F, Inglada J and Gancarski P. 2012. Satellite image time series analysis under time warping. IEEE Transactions on Geoscience and Remote Sensing, 50(8): 3081-3095 [DOI: 10.1109/TGRS.2011.2179050]

Rußwurm M and Körner M. 2017. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu: IEEE: 1496-1504 [DOI: 10.1109/CVPRW.2017.193]

Rußwurm M and Körner M. 2018. Multi-temporal land cover classification with sequential recurrent encoders. ISPRS International Journal of Geo-Information, 7(4): 129 [DOI: 10.3390/ijgi7040129]

Rußwurm M., & Körner M. 2020. Self-attention for raw optical satellite time series classification. ISPRS journal of photogrammetry and remote sensing, 169, 421-435 [DOI: 10.1016/j.isprsjprs.2020.06.006]

Sutskever I, Vinyals O and Le Q V. 2014. Sequence to sequence learning with neural networks//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press: 3104-3112

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.: 6000-6010

Zhang Z, Tang P and Duan R B. 2015. Dynamic time warping under pointwise shape context. Information Sciences, 315: 88-101 [DOI: 10.1016/j.ins.2015.04.007]

Zhang Z, Tang P, Huo L Z and Zhou Z G. 2014. MODIS NDVI time series clustering under dynamic time warping. International Journal of Wavelets, Multiresolution and Information Processing, 12(5): 1461011 [DOI: 10.1142/S0219691314610116]

Zhang Z, Tavenard R, Bailly A, Tang X T, Tang P and Corpetti T. 2017. Dynamic time warping under limited warping path length. Information Sciences, 393: 91-107 [DOI: 10.1016/j.ins.2017.02.018]

Zhong L H, Hu L N and Zhou H. 2019. Deep learning based multi-temporal crop classification. Remote Sensing of Environment, 221: 430-443 [DOI: 10.1016/j.rse.2018.11.032]

Time series classification of remote sensing data based on temporal self-attention mechanism

ZHANG Weixiong[1,2], TANG Ping[1], ZHANG Zheng[1]

1.Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China;
2.University of Chinese Academy of Sciences, School of Electronic, Electrical and Communication Engineering, Beijing 100049, China

Abstract: With the rapid development of remote sensing technology, the continuous accumulation of remote sensing time series data provides an important data support for studying land cover classification. Extracting classificational discriminative features from remote sensing time series data by using deep learning methods has become a hot research topic. Deep learning methods require a large number of training data, but sample imbalance prevents the commonly used recurrent and convolutional networks from achieving high accuracies in categories that have a small number of samples. To address this problem, this paper introduces the self-attention mechanism originating from the field of natural language processing to the classification of multispectral remote sensing time series data with the aim of extracting deep temporal features at a global scale. This mechanism differs from recurrent networks, which extract temporal features by using the previous time information along the temporal dimension, and from convolutional networks, which extract temporal features at the local time neighborhood.

We construct a new feature extraction network based on the transformer encoder, which initially employs the self-attention mechanism in natural language processing, and then compare this neatwork with the long- and short-term-memory-based feature extraction network and temporal-convolution-neural-network-based feature extraction network to evaluate the effectiveness of the self-attention-based method in improving the classification accuracy of small-sample categories. To achieve a fair comparison, we adopt a generic classification framework consisting of data input, feature extraction network, classifier, and classification output, and we use different models with various hyperparameters as feature extraction networks. We then evaluate the classification performance of different methods on the TiSeLaC public multispectral remote sensing time series dataset by using per-class accuracy, overall accuracy (OA), and mean intersection over union (mIoU) as metrics.

To obtain a proper measure of different methods, we choose the top three mIoU hyperparameter settings for each model and then calculate the average metrics as the final result. Results show that the self-attention-based network outperforms both the recurrent and convolutional networks. This network achieves a 92.98% OA and 80.60% mIoU, which are 1.25% and 1.32% higher than those achieved by the recurrent and convolutional networks, respectively. In terms of per-class accuracy, while the self-attention-based network achieves equivalent accuracies with differences of less than 0.74% in the large-sample categories compared with the recurrent and convolutional networks, the proposed network can significantly improve classification accuracies in small-sample categories by large margins ranging from 2.47% to 5.41%.

This paper introduces the self-attention mechanism to the classification of multispectral remote sensing time series data to address the problem of low classification accuracy in small-sample categories caused by sample imbalance. We construct a new temporal feature extraction network based on the self-attention mechanism to globally extract temporal features from time series and design a set of objective comparison experiments. Experiment results show that by globally extracting temporal features from time series, instead of using previous time information (as in the case of recurrent networks) and focusing on the local time neighborhood (as in the case of convolutional networks), the self-attention-based network achieves the same accuracy in majority-sample categories and effectively improves the accuracy in small-sample categories. Therefore, the self-attention-based network can play an important role in the future classification of remote sensing time series, and further research on this network is critical.

Key words: self-attention mechanism, deep learning, remote sensing time series, land cover classification, imbalance of samples

Supported by National Natural Science Foundation of China (No. 41701399, 41971396, 41701397)