引用格式：廖健文，杨盈昀，卢玥. 基于稀疏Transformer的长短时序关联动作识别算法[J]. 中国传媒大学学报（自然科学版），2023，30(06)：56-63.
文章编号：1673-4793(2023)06-0056-08

基于稀疏Transformer的长短时序关联动作识别算法

廖健文，杨盈昀*，卢玥
（中国传媒大学信息与通信工程学院，北京 100024）

摘要：针对主流的视频动作识别算法对时序信息的挖掘不充分，而Transformer能够更好地处理长序列和全局依赖性问题，本文将3DCNN和Transformer结合起来，提出了基于稀疏Transformer的长短时序关联动作识别算法，从而实现对视频的全局时序信息进行建模。该算法提取预训练视频模型各个片段特征，嵌入视频特征聚类模块降低输入特征的潜在噪声，并利用基于稀疏自注意力的Transformer长短时序关联模块，引入稀疏掩码矩阵，对相似度矩阵进行掩码操作，抑制较小的注意力权重，选择性地保留重要的长短时序信息，提高模型对全局上下文信息的注意力集中程度。本文在UCF101和HMDB51数据集上进行了大量的实验，验证了本文算法的有效性，在参数量和计算复杂度较小的情况下准确率高于同类权威算法。

关键词：深度学习；动作识别；稀疏Transformer；R3D-18
中图分类号：TP183　**文献标识码**：A

Sparse transformer-based algorithm for long-short temporal association action recognition

LIAO Jianwen, YANG Yingyun*, LU Yue
(School of Information and Communication Engineering, Communication University of China, Beijing 100024, China)

Abstract: Mainstream video action recognition algorithms often lack sufficient exploitation of temporal information, while Transformer excels at handling long sequences and global dependency issues. In this paper 3D Convolutional Neural Networks(3DCNN) and Transformer were combined to propose a sparse Transformer-based long-short temporal association action recognition algorithm, so as to realize the modeling of global temporal information of video. The algorithm used a pre-trained model to extract clip features, embedded a video feature clustering module to reduce the potential noise of the input features, and used a Transformer long-short temporal association module based on sparse self-attentiveness which introduced a sparse mask matrix masking operations on the similarity matrix to suppress smaller attention weights, selectively retained important long-short temporal information, and improved the model's attention concentration on global contextual information. The experimental results verify the effectiveness of the proposed algorithm, showing the model can achieve higher accuracy compared to state-of-the-art approaches on the UCF101 and HMDB51 datasets with a small number of parameters and computational complexity.

Key words: deep learning; action recognition; sparse transformer; R3D-18

作者简介（*为通信作者）：廖健文（2000-），女，硕士研究生，主要从事视频理解研究。Email：liaojianwen@cuc.edu.cn；杨盈昀（1969-），女，博士，教授，主要从事智能视音频分析与处理。Email：yingyun6903@163.com；卢悦玥（1998-），女，硕士研究生，主要从事视频理解研究。Email：luyue@cuc.edu.cn

1 引言

近年来，随着移动互联网技术的发展，以及社交娱乐平台软件的普及，网络视频行业正面临着机遇与挑战。短视频与直播已成为当前互联网主流的内容表现形式，这种趋势推动了海量视频数据的涌现。然而，如何有效地分析、处理、分类和管理海量视频数据成为研究人员面临着的一个重大挑战。视频理解算法应运而生，其核心目标在于自动提取视频数据中的特征，并据此推断出视频中所展现的特定动作或行为。这类算法依赖于对视频数据的理解和分析，包括对人体姿势、关键动作、运动模式及其时序特征的捕获。在动作识别算法中，重要的挑战之一是有效地捕捉和表示视频中的运动特征，以便进行有效的分类和识别。动作识别算法将视频数据与动作标签联系在一起，既处理每个视频帧中的图像内容，又挖掘视频帧之间的时序信息，从而提取出视频序列中的人类动作特征，获得更加准确和鲁棒的识别结果。

随着深度学习技术的迅猛发展，研究人员也逐步转向研究基于深度学习的人体动作识别算法。卷积神经网络（Convolutional Neural Networks, CNN）具有局部感受野、共享权值和空间子采样的优势，但并不能捕捉全局特征。基于长短期记忆网络（Long Short-Term Memory, LSTM）的动作识别算法在CNN之后添加 LSTM 或门控循环单元（Gated Recurrent Unit, GRU），来提取视频的时序信息。但当处理长序列数据时，LSTM层数或GRU个数的增多让网络模型更容易产生梯度爆炸、难以优化的问题。相较而言，Transformer模型没有使用递归或循环机制，而是使用了自注意力机制来实现序列数据的处理，在关注和建模序列中的全局上下文信息，以及并行计算等方面都表现良好。因此，在卷积神经网络提取特征的基础上结合Transformer模型，能够兼具捕捉局部和全局特征的能力，对长短时序信息进行关联，可以更好地处理图像和序列数据，也可以提高准确率、训练和推理的速度。

然而，Transformer模型在处理视频数据时随着序列数据长度的增加，模型训练和推理时所需的内存和计算复杂度呈现二次方增长，这也会导致模型难以训练和收敛，限制了算法的精度上限，存在着计算量大、浪费计算资源等问题。针对上述问题，本文对Transformer模型进行优化和改进，提出了视频特征聚类模块和自注意力稀疏化方法。

2 相关研究

主流的视频动作识别算法主要分为基于双流结构的动作识别、基于长短时记忆网络的动作识别和基于三维卷积神经网络（3D Convolutional Neural Networks, 3DCNN）[1]的动作识别。

基于3DCNN的动作识别算法能够捕获多个连续视频帧中的运动信息，将多个视频帧的多通道信息堆叠在一起构成一个“立方体”，将其与一个3D核进行卷积，完成三维卷积操作，提取到融合的时空特征。可以直接从时间、通道和空间这三个维度的视频数据中学习到时空运动和空间特征，并且可以在不计算大量光流的情况下取得良好的结果。

卷积三维网络（Convolutional 3D, C3D）[2]证明与二维卷积神经网络相比，三维卷积和三维池化操作能够在建模空间信息的同时也很好地建模时间信息，以及3 × 3 × 3的三维卷积核能够提取出最为紧凑和显著的时空特征。Hara等人[3]将残差网络（ResNet）的思想引入3DCNN中，提出3D ResNet（3D Residual Network）网络模型，将ResNet中的2D卷积层替换成3D卷积层，大幅减少了参数量（Params）和每秒浮点运算次数（Floating-point Operations Per second，FLOPs）。伪三维残差网络（Pseudo-3D Residual Net, P3D）[4]将3 × 3 × 3的时空卷积核进行分离，变成1 × 3 × 3空间卷积核和3 × 1 × 1时间卷积核，进一步减少参数量和FLOPs。

Tran等人[5]总结了上述的改进方法，对3D ResNet网络模型进行小幅度的修改，提出R3D（Residual 3D Network）和R(2+1)D（Residual (2+1)D Network）网络模型，并证明$t \times d \times d$的时空卷积核可以分解成一个$1 \times d \times d$的空间卷积核和若干个$t \times 1 \times 1$时间卷积核。R3D网络模型的参数量是C3D网络模型的一半，运行速度也比C3D网络模型快了一倍。

双流膨胀三维卷积网络（Inflated 3D ConvNets, I3D）[6]使用三维卷积和三维池化来替代Inception网络中的二维卷积和二维池化，允许输入更大时空分辨率的数据。分离三维卷积网络（Separate 3D, S3D）[7]在I3D的基础上将时空卷积分离，能够更好地学习和提取到更抽象的时空特征。

上述这些网络的输入都是剪辑后的16帧或多帧图像，而不是从完整的视频中学习，从而忽略了视频中的远程时空依赖性。基于3DCNN的人体动作识别算法，虽然可以从时间、通道和空间三个维度来学习输入视频

数据中包含的时序运动和空间特征,能够在不计算大量光流数据的情况下依旧取得不错的效果,但其仍然存在着时空特征提取能力不够高效、参数和计算量过多、需要大规模数据集和网络优化困难等诸多问题。

许多研究人员尝试将Transformer模型运用到其他计算机视觉任务中,如时间信息预测、问答系统、多模态匹配、推荐算法以及各类计算机视觉问题。Video Transformer(ViT)[8]对输入图像进行切片,将切片结果作为Transformer模型的输入序列,在图像分类任务上达到了很好的效果。因此,Transformer模型也可探索推广到视频理解算法上来,比如Video Vision Transformer(ViViT)[9]。在人体动作识别算法中,Jin等人[10]引入骨髓网络和序列权重模块来对Transformer模型进行改造升级,将其连接在R(2+1)D网络模型之后,利用改进的Transformer模型提取更加抽象和高级的时空特征,进行时序建模。Kalfaoglu等人[11]利用BERT模型来代替3DCNN中的全局池化层,更好地提取时间信息,提高了多种3DCNN的动作识别性能。

为了能够轻量化使用Transformer模型,Linear Transformer[12]将传统的*Softmax*注意力转变为基于特征映射的线性点积注意力,大大节省了内存使用率及时间复杂度。Sparse Transformer[13]对多头注意力进行稀疏化,计算多种局部的自注意力,既加强了局部信息的紧密性,又降低了计算复杂度。

与上述思路不同,本文提出了视频特征聚类模块和自注意力稀疏化方法,对Transformer模型进行优化和改进。视频特征聚类模块能够对输入的片段特征的个数进行限制,筛选掉一部分会降低模型推理精度的片段特征;自注意力稀疏化方法对数值小的注意力权重进行抑制,选择性地保留重要的长短时序信息,提高模型对全局上下文信息的注意力集中程度,建立长短时序信息的依赖关系。本文将这两个模块应用到原始Transformer模型中,并在UCF101[14]和HMDB51[15]数据集上利用R3D-18预训练模型提取视频高维时空特征进行了训练和测试。

3 模型框架

如图1所示,本文提出的基于稀疏Transformer的长短时间关联动作识别算法模型主要分为四个部分:基于R3D-18预训练模型的特征提取模块通过加载在Kinetics-700数据集[16]上预训练的R3D-18模型来对分片后的各

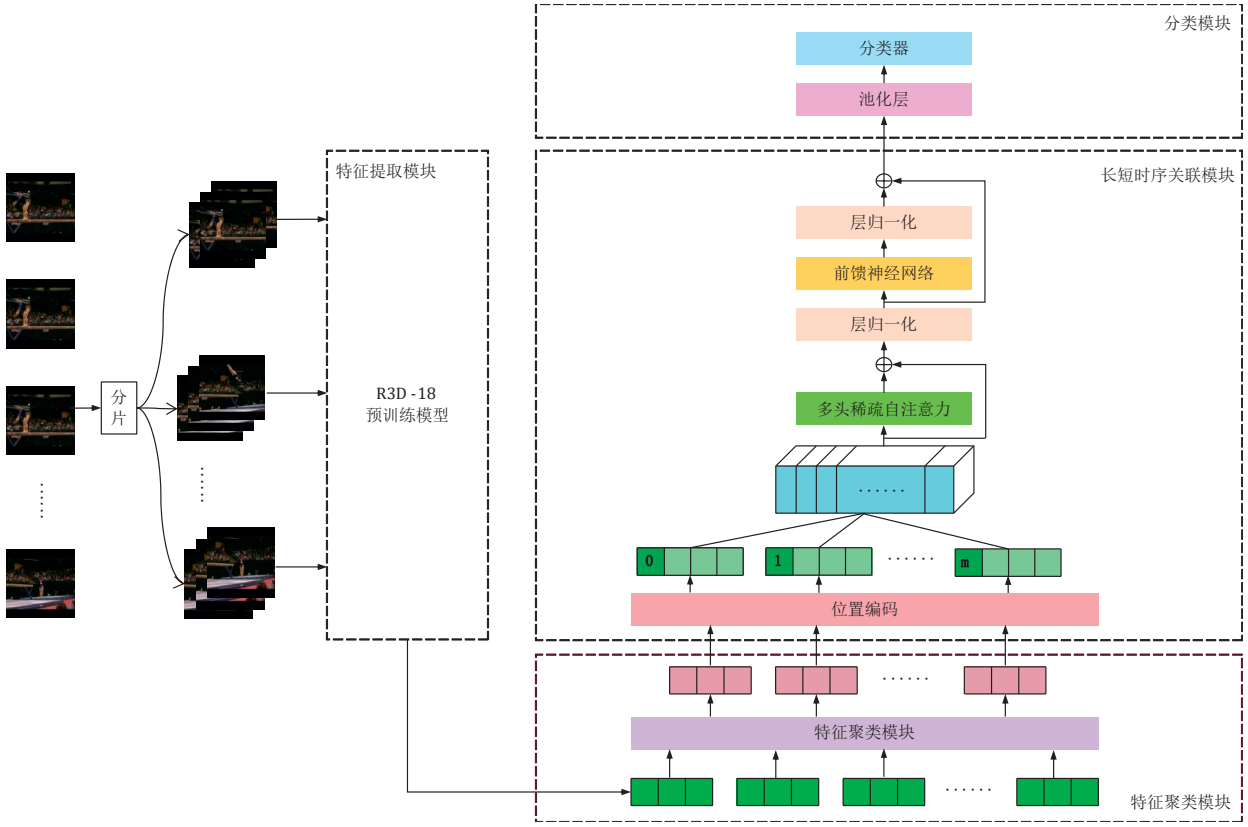


图1 基于稀疏Transformer的长短时序关联算法模型

个片段进行时空特征的提取；视频特征聚类模块通过计算基准中心特征和比较各片段特征距其的余弦距离，来限制输入的片段特征个数，剔除潜在的数据噪声，减少模型所需的计算量和内存；基于稀疏自注意力的Transformer长短时序关联模块在原始Transformer模型编码器的基础上将多头自注意力模块换成多头稀疏自注意力模块，通过对数值较低的注意力权重进行置0抑制，保留数值较大的注意力权重，减少参与运算的数据，提高自注意力的集中程度；分类模块由平均池化层和分类器组成，分类器包括一个全连接层和*Softmax*函数，得到最终的动作分类结果。

3.1 视频特征聚类模块

类似R3D-18这样的分类网络模型在提取特征时，提取得到的高维特征通常在高维特征空间中呈现簇状聚集的现象。一般来说，网络模型分类能力越高，同一类特征聚集得更加集中，不同类别之间的特征也会相距较远，形成一个比较明显的决策超平面。网络模型在决策超平面中通过不同类别特征之间的决策边线来实现分类任务。然而，R3D-18网络模型的分类准确率并不高，提取得到的同一类特征中存在一些偏离聚集中心的离散特征，这些离散特征可能反而和其他类的特征相近，容易影响后续Transformer模型的训练和推理效果。把这些离散特征看作噪声，使用特征筛选方法来筛除这些离散特征，只保留最为聚集的特征，调整得到决策超平面，既可以减少输入的片段特征个数，降低整个模型的计算量和所需内存，又可以提高后续Transformer模型的准确率。

基于上述问题分析，本文提出视频特征聚类模块，根据每一类的所有片段特征选定基准中心特征，将这一类的所有片段特征和该类的基准中心特征进行余弦距离的计算。余弦距离的数值大，则说明该片段特征与基准中心特征距离较远，即该片段特征比较离群；余弦距离的数值小，则说明该片段特征离基准中心特征比较近，即该片段特征更靠近该类的中心。最后根据余弦距离的大小来对片段特征进行筛选，筛选掉余弦距离较大的片段特征，保留余弦距离较小的片段特征作为模型的输入。对于整个视频，加载R3D-18预训练网络模型来提取得到片段特征，得到整个视频序列特征 $X=\{F_1,F_2,\cdots,F_n\}$，其中 $F_i\in\mathbb{R}^{1\times 512}$ 表示为第 i 个片段特征，n 表示为整个视频序列有 n 个片段特征，即 $X\in\mathbb{R}^{n\times 512}$。在选定基准中心特征 $C\in\mathbb{R}^{1\times 512}$ 时，对这 n 个片段特征沿着各维度取平均值，如式(1)所示：

$$C=Average(X)=Average\{F_1,F_2,\cdots,F_n\} \tag{1}$$

计算得到基准中心特征后，为了衡量各个片段特征与其的距离，计算片段特征 $\{F_1,F_2,\cdots,F_n\}$ 和基准中心特征 C 之间的余弦距离，公式如式(2)：

$$dist_i=1-\frac{F_i\cdot C}{\|F_i\|\|C\|} \tag{2}$$

其中，$\|F_i\|$ 和 $\|C\|$ 分别是第 i 个片段特征 F_i 和基准中心特征 C 的模长。根据距离列表 $D=\{dist_1,dist_2,\cdots,dist_n\}\in\mathbb{R}^{1\times n}$，将余弦距离较大的片段特征筛出，仅保留剩下的片段特征 $X'=\{F'_1,F'_2,\cdots,F'_m\}$。

为了适应不同的数据集和网络模型，本文提出的视频特征聚类模块针对输入片段特征的个数设置了阈值 L 和过滤因子 r，而不是设置一个特定的余弦距离阈值，因为特定的余弦距离阈值是一个相当敏感的超参数，具体处理流程如图2所示。

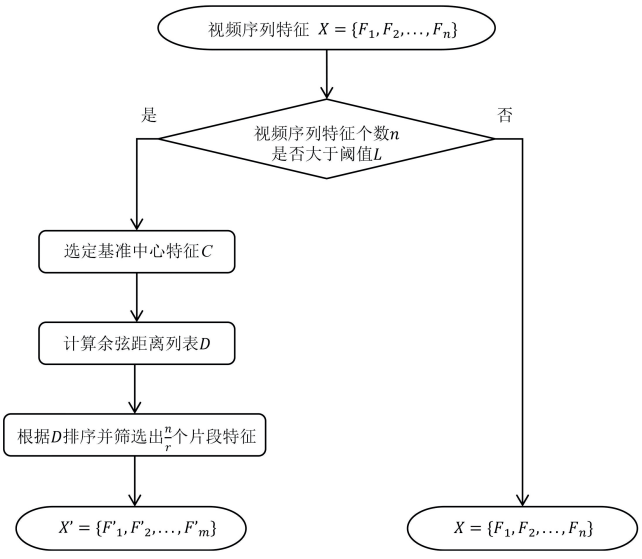


图2 视频特征聚类流程

当原始输入的片段特征个数 n 不大于阈值 L 时，说明此时序列特征 X 的长度 n 是模型可全部接受的长度，不需要进行特征筛选操作；当序列特征 X 的长度大于阈值 L 时，说明此时序列特征 X 属于长视频的范畴，需要进行特征筛选。在排序和筛选时，过滤因子 r 限制筛选掉的片段特征个数，将余弦距离最大的对应的 $\frac{n}{r}$ 个片段抛弃，保留剩下的 m 个片段特征。m 的计算如式(3)：

$$m = n - \frac{n}{r} \tag{3}$$

3.2 基于稀疏自注意力的Transformer长短时序关联模块

Transformer模型利用自注意力机制对整个序列的所有片段特征进行时序建模,计算出注意力权重,当注意力权重较小时,说明模型不太关注所对应的这两个片段特征间的时序关系,使得模型在训练和推理的过程中产生大量的"无用的"的计算。那么在后续计算中可以忽略这些片段特征的注意力权重的计算,对更加重要的长短时序信息进行关联,将计算资源更加集中于注意力权重大的那些片段特征。同时,无论多小的注意力权重,都对接下来的注意力权重计算结果产生影响,进而影响模型对动作识别的准确率。

假设在通过视频特征聚类模块后,输入稀疏Transformer模块的序列特征为 $X' = \{F'_1, F'_2, \cdots, F'_m\}$。在自注意力机制中,先分别将查询矩阵 W_Q、键矩阵 W_K 和值矩阵 W_V 转换为查询向量 $Q \in \mathbb{R}^{m \times 512}$、键向量 $K \in \mathbb{R}^{m \times 512}$ 和值向量 $V \in \mathbb{R}^{m \times 512}$,后通过一系列计算得到自注意力模块的输出 $Y \in \mathbb{R}^{m \times 512}$,具体计算过程如式(4)所示:

$$\boldsymbol{Y} = Attention(Q,K,V) = Soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

相似度矩阵 S 是上式中的 $\frac{QK^T}{\sqrt{d_k}}$,并且 $S \in \mathbb{R}^{m \times m}$,$d_k$ 为隐藏层的维度。注意力权重矩阵 A 是上式中的 $Soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)$,并且 $A \in \mathbb{R}^{m \times m}$。相似度得分 $S_{i,j}$ 的数值比较大,则经过 $Soft\max$ 函数后得到的注意力权重数值也会比较大,反映出第 i 个片段特征 F'_i 与第 j 个片段特征 F'_j 之间有着比较密切的联系,是值得模型关注的。对于相似度得分 S_i 而言,其元素之和 $\sum_{j=1}^{m} S_{i,j}$ 则反映了第 i 个片段特征 F'_i 在输入的整个序列特征 X' 中的重要程度。

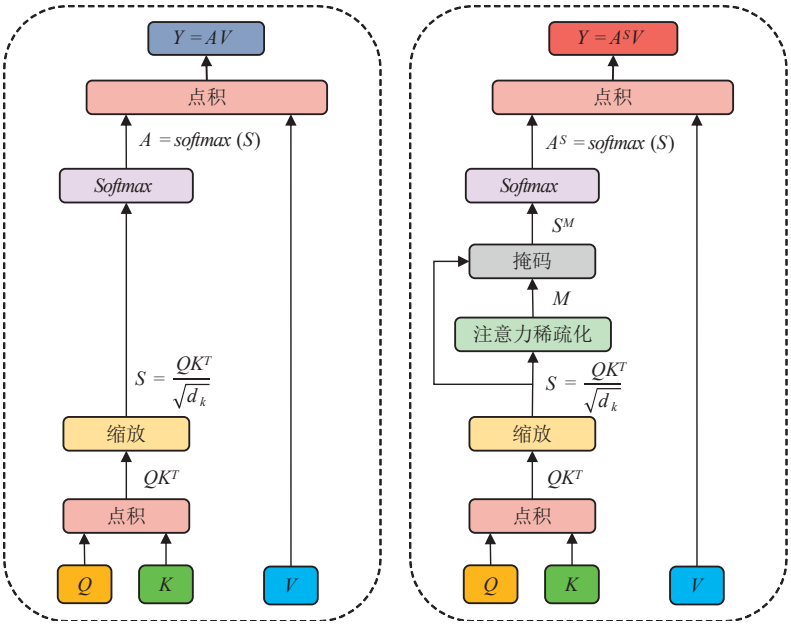
本文所提出的基于稀疏自注意力的Transformer长短时序关联模块先放眼于整个视频序列特征,保留对于整个视频序列而言比较重要的片段特征,即根据 $\sum_{j=1}^{m} S_{i,j}$ 的排序结果,保留和最大的 $\frac{m}{2}$ 行相似度得分。之后,为了保留尽可能多的时序信息,并没有立刻丢弃其余的 $\frac{m}{2}$ 行相似度得分,而是在其余的 $\frac{m}{2}$ 行相似度得分中,对每行相似度得分中的元素 $S_{i,j}$ 进行排序,保留每行中最大的 $\frac{m}{4}$ 个相似度得分,抑制剩下的 $\frac{3m}{4}$ 个相似度得分,使其对应位置的注意力权重为0,不参与接下来的运算,使模型关注着拥有紧密关系的片段特征间的时序信息。最终,得到稀疏注意力权重矩阵 $A^s \in \mathbb{R}^{m \times m}$。因此,本文提出稀疏注意力机制的计算过程表示为式(5)。

$$\begin{aligned} Y &= SparseAttention(Q,K,V) \\ &= Softmax\left[Sparse\left(\frac{QK^T}{\sqrt{d_k}}\right)\right]V \end{aligned} \tag{5}$$

如图3所示,为了避免强行置0影响整个模型的梯度更新、反向传播,本文引入了一个稀疏掩码矩阵 $M \in \mathbb{R}^{m \times m}$,来将上述的自注意力稀疏化过程转换为函数运算的过程。首先,根据各个需要抑制和丢弃的注意力权重的位置,将稀疏掩码矩阵 M 中所对应位置的数值置为1,而对于其他的需要继续保留的注意力权重,稀疏掩码矩阵 M 中对应位置的数值是0。之后,利用生成的稀疏掩码矩阵 M 对相似度矩阵 S 进行掩码操作实现数值变换,用一个无穷小的值来填充相似度矩阵 S 中与稀疏掩码矩阵 M 中值为1位置相对应的元素,得到了掩码相似度矩阵 $S^M \in \mathbb{R}^{m \times m}$。当一个无穷小的值与其他数值一起经过 $Soft\max$ 函数时,无穷小的数值会被转换为0。最后,对掩码相似度矩阵 S^M 进行 $Softmax$ 函数的归一化操作,无穷小对应位置的数值就被设置为了0,得到稀疏注意力权重矩阵 $A^S \in \mathbb{R}^{m \times m}$,而其他数值继续进行了归一化操作,重要信息继续参与后续的计算中,模型会更加重点关注这些被保留下来的时序信息。

以相似度矩阵 $S \in \mathbb{R}^{4 \times 4}$ 为一个简单的例子来更加清楚地说明上述的自注意力稀疏化计算过程,图4展示了具体的计算过程。

在稀疏注意力权重矩阵 A^s 与值向量 V 相乘后,模型会保持着注意力权重的原本分配方式,该片段特征会加权融合到经自注意力模块提取到的全局时序信息,对该片段特征与其他所有的片段特征之间的潜在长短时序信息进行关联;而对于在整个序列中不那么重要的片段特征,模型只会将注意力权重集中于与其有着紧密时空关系的某些片段特征上,忽略掉不重要的时序信息,关联最重要的长短时序信息。



(a) 原始Transformer的自注意力机制 (b) 改进后的稀疏自注意力机制

图3 改进前后自注意力机制的流程对比图

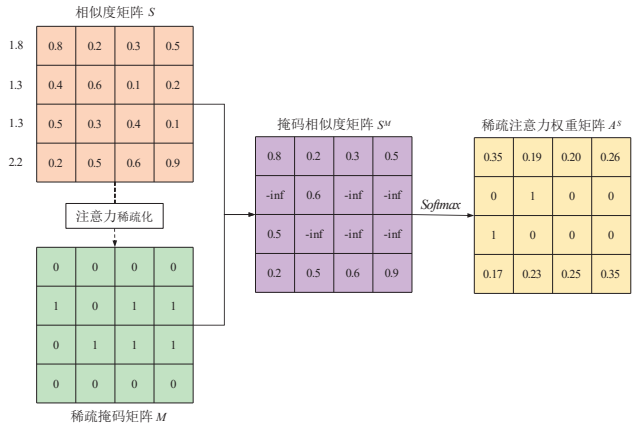


图4 自注意力稀疏化计算过程的具体例子

4 实验结果与分析

4.1 数据集

本文提出的模型在UCF101和HMDB51数据集上进行了验证。UCF101数据集中的视频来源于YouTube视频网站上由用户自己上传的视频。该数据集共101个动作类别，平均每个动作类别有约100个视频，一共有13320个视频。HMDB51数据集中大部分视频是从电影中剪辑好的片段，只有少数视频来源于其他公开数据集和在线视频网站。该数据集共51个动作类别，平均每个动作类别有约100个视频，共有6766个视频。

4.2 实验环境设置

对于整体模型训练，先对训练集中每个视频进行分片，每16个相邻的视频帧为一个片段，再对每个片段进行逐帧抽取，利用尺度抖动、水平翻转这两种数据增强方法进一步增加训练数据的规模，所有RGB图像帧的大小固定为112 × 112，之后把视频的所有片段数据作为模型的输入。提取片段特征的预训练模型是在Kinetics-700数据集上预训练后的R3D-18模型[17]。整个网络模型的初始学习率被设置为0.01。本实验使用了随机梯度下降法来优化整个模型，即使用SGD优化器，SGD优化器的动量为0.9，并且学习率在每10个Epoch结束后都会衰减到当前的0.9，整个实验总共经过200个Epoch的训练。网络模型计算损失的公式是交叉熵函数。衡量指标是网络模型输出的Top-1识别准确率。

在模型的验证与测试上，采用与训练时相同的方式来对测试集视频数据进行预处理。本文使用一张NVIDIA GeForce RTX 3080显卡来加速模型的训练和测试。

在模型的超参数方面，视频特征聚类模块中有个数阈值和筛选因子。在实验中，个数阈值被设置为50，筛选因子被设置为2。

4.3 实验结果对比与分析

在本节中，我们将所提出的模型与其他研究方法进行了比较。在表1中展示了本文方法及以往研究方法在UCF101和HMDB51数据集上的动作识别Top-1准确率。从表中可以看出，在UCF101数据集上，本文的

模型胜过了大多数先前的研究方法,并达到了97.41%的准确率。这说明本文提出的视频特征聚类模块和基于稀疏注意力的Transformer模块在建模时序信息和提取视频全局时空特征方面有着更好的能力。虽然本文提出的算法只比未经预训练的VidTr-L模型高出0.71%,但VidTr-L模型需要输入32帧视频图像,并且FLOPs为351G,是本文算法模型FLOPs的8.59倍。

在HMDB51数据集上,相较于TS-LSTM和C^2LSTM,本文提出的算法模型表现也非常出色。但本文算法模型的准确率比D3D仅高了0.09%,这是因为D3D引入知识蒸馏思想,从教师模型中获取光流运动信息,即在训练时引入了光流数据,不仅增加了不同模态的数据,而且增加了参数量和计算复杂度。

然而,在UCF101和HMDB51数据集上,本文算法模型比R(2+1)D+BERT(32f)算法模型分别低了1.24%和5.2%的识别准确率。R(2+1)D+BERT(32f)算法:一方面增加了输入的片段帧数,从16帧增加了一倍到32帧,以及增加了R(2+1)D网络的深度,使用了R(2+1)D-34模型;另一方面该算法利用BERT来代替3DCNN模型最后的时间全局平均池化层,建模时序信息。整个算法模型需要66.67M的参数量和152.97G的FLOPs,参数和计算量巨大,且结构十分复杂。

表1 本文算法与其他方法在不同数据集上的实验结果

方法	预训练数据	UCF101准确率(%)	HMDB51准确率(%)	参数量	FLOP
TS-LSTM[18]	–	94.10	69.00	–	–
C^2LSTM[19]	–	92.80	61.30	–	–
TSM[20]	Kinetics-400	95.90	73.50	**23.4M**	65G
VideoMAE[21]	Kinetics-400	96.10	73.30	87M	180 × 5 × 3G
MVFNet[22]	Kinetics-400	96.60	75.70	–	66G
VidTr-L[23]	–	96.70	74.40	–	351G
D3D[24]	Kinetics-400	97.00	78.70	–	–
R(2+1)D+Impoved Transformer[10]	Kinetics-400	97.18	–	32.50M	42.61G
R(2+1)D+BERT(32f)[11]	IG65M	**98.65**	**83.99**	66.67M	152.97G
Ours	Kinetics-700	97.41	78.79	36.69M	**40.87G**

4.4 消融实验

在本节中,我们进行了消融实验来探索每个模块对整个算法模型的贡献。消融实验结果如表2所示,在UCF101数据集上,预训练模型结合原始Transformer模块可以达到97.04%准确率,反映了Transformer模块可以建模输入片段特征间潜在的长短时序信息,提高动作识别的准确率。

在参数量和计算复杂度方面,嵌入原始Transformer模块后,本文算法模型增加了3.46 M的参数量和0.02 G的FLOPs,分别需要36.69 M的参数量和40.87 GFLOPs。当嵌入其他两个模块后,本文算法模型的参数量为36.69 M,FLOPs为40.87 G,均没有发生明显的增加。

对于个数阈值L和筛选因子r这两个超参数,本文对不同超参设置下的网络模型进行了消融实验,结果如表3所示。

可以看到当个数阈值L设置为50且筛选因子r设置为2时,动作识别准确率最高。

表2 UCF101数据集上关于各模块有效性的消融实验结果

方法			准确率(%)	参数量	FLOPs
Transformer	视频特征聚类	稀疏自注意力			
×	×	×	95.45	**33.23 M**	**40.85 G**
√	×	×	97.04	36.69 M	40.87 G
√	√	×	97.15	36.69 M	40.87 G
√	×	√	97.22	36.69 M	40.87 G
√	√	√	**97.41**	36.69 M	40.87 G

表3 UCF101数据集上关于L,r的消融实验结果

(L,r)	准确率(%)
(40, 2)	97.19
(40, 3)	97.26
(50, 2)	**97.41**
(50, 3)	97.26

5 结论

为了补充3DCNN所缺乏的时序信息,进一步地降低模型计算量,加强模型对注意力权重的关注程

度，本文提出了基于稀疏Transformer的长短时序关联动作识别算法模型，加载预训练模型之后利用稀疏Transformer模型来建模长短时序信息，捕捉视频序列的全局和局部时空特征。本文提出的基于稀疏Transformer的长短时序关联算法模型将输入视频片段切分成片段，利用R3D-18预训练模型来处理视频数据，在UCF101和HMDB51数据集上，进行了充分的实验，验证了算法的有效性。

参考文献（References）：

[1] Ji S，Xu W，Yang M，et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2012，35(1)：221-231.

[2] Tran D，Bourdev L，Fergus R，et al. Learning spatiotemporal features with 3D convolutional networks[C]// Proceeding of the IEEE International Conference on Computer Vision，2015：4489-4497.

[3] Hara K，Kataoka H，Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition [C]// Proceeding of the IEEE International Conference on Computer Vision Workshops，2017：3154-3160.

[4] Qiu Z，Yao T，Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]// Proceeding of the IEEE International Conference on Computer Vision，2017：5533-5541.

[5] Tran D，Wang H，Torresani L，et al. A closer look at spatiotemporal convolutions for action recognition [C]// Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition，2018：6450-6459.

[6] Carreira J，Zisserman A. Quo vadis，action recognition? a new model and the kinetics dataset[C]// Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition，2017：6299-6308.

[7] Xie S，Sun C，Huang J，et al. Rethinking spatiotemporal feature learning：speed-accuracy trade-offs in video classification[C]// Proceeding of the European Conference on Computer Vision（ECCV），2018：305-321.

[8] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16x16 words：transformers for image recognition at scale[DB/OL]. arXiv：2010.11929，2020.

[9] Arnab A，Dehghani M，Heigold G，et al. ViViT：a video vision transformer [C]// 2021 IEEE/CVF International Conference on Computer Vision（ICCV），2021：6816-6826.

[10] Jin H，Yang J，Zhang S. Efficient action eecognition with introducing R(2+1)D convolution to improved transformer [C]// 4th International Conference on Information Communication and Signal Processing（ICICSP），2021.

[11] Kalfaoglu M E，Kalkan S，Alatan A A. Late temporal modeling in 3D CNN architectures with BERT for action recognition[C]// Proceeding of Computer Vision-ECCV 2020 Workshops，2020：731-747.

[12] Katharopoulos A，Vyas A，Pappas N，et al. Transformers are RNNs：fast autoregressive transformers with linear attention [C]// International Conference on Machine Learning，2020：5156-5165.

[13] Child R，Gray S，Radford A，et al. Generating long sequences with sparse transformers[DB/OL]. arXiv：1904.10509，2019.

[14] Soomro K，Zamir A R，Shah M. UCF101：a dataset of 101 human actions classes from videos in the wild[DB/OL]. arXiv：1212.0402，2012.

[15] Kuehne H，Jhuang H，Garrote E，et al. HMDB：a large video database for human motion recognition[C]// International Conference on Computer Vision（ICCV），2011.

[16] Carreira J，Noland E，Hillier C，et al. A short note on the kinetics-700 human action dataset[DB/OL]. arXiv：1907.06987，2019.

[17] Hirokatsu K，Tenga W，Kensho H，et al. Would mega-scale datasets further enhance spatiotemporal 3D CNNs[DB/OL]. arXiv：2004.04968，2020.

[18] Ma C Y，Chen M H，Kira Z，et al. TS-LSTM and temporal-inception：exploiting spatiotemporal dynamics for activity recognition[DB/OL]. arXiv：1703.10667，2017.

[19] Majd M，Safabakhsh R. Correlational convolutional LSTM for human action recognition[J]. Neurocomputing，2020，396：224-229.

[20] Lin J，Gan C，Han S. TSM：temporal shift module for efficient video understanding[C]// 2019 IEEE/CVF International Conference on Computer Vision，2019，7082-7092.

[21] Tong Z，Song Y，Wang J，et al. VideoMAE：masked autoencoders are data-efficient learners for self-supervised video pre-training [J]. Advances in Neural Information Processing Systems，2022，35：10078-10093.

[22] Wu W，He D，Lin T，et al. MVFNet：multi-view fusion network for efficient video recognition[C]//Proceeding of the AAAI Conference on Artificial Intelligence，2021，35(4)：2943-2951.

[23] Zhang Y，Li X，Liu C，et al. VidTr：video transformer without convolutions[DB/OL]. arXiv：2104.11746，2021.

[24] Stroud J C，Ross D A，Sun C，et al. D3D：distilled 3D networks for video action recognition [C]// 2020 IEEE Winter Conference on Applications of Computer Vision（WACV），2020：614-623.

编辑：赵志军