基于骨架动作识别的协作卷积Transformer网络

石跃祥 朱茂清*

(湘潭大学计算机学院网络空间安全学院 湘潭 411105)

摘 要：近年来，基于骨架的人体动作识别任务因骨架数据的鲁棒性和泛化能力而受到了广泛关注。其中，将人体骨骼建模为时空图的图卷积网络取得了显著的性能。然而图卷积主要通过一系列3D卷积来学习长期交互联系，这种联系偏向于局部并且受到卷积核大小的限制，无法有效地捕获远程依赖关系。该文提出一种协作卷积Transformer网络(Co-ConvT)，通过引入Transformer中的自注意力机制建立远程依赖关系，并将其与图卷积神经网络(GCNs)相结合进行动作识别，使模型既能通过图卷积神经网络提取局部信息，也能通过Transformer捕获丰富的远程依赖项。另外，Transformer的自注意力机制在像素级进行计算，因此产生了极大的计算代价，该模型通过将整个网络分为两个阶段，第1阶段使用纯卷积来提取浅层空间特征，第2阶段使用所提出的ConvT块捕获高层语义信息，降低了计算复杂度。此外，原始Transformer中的线性嵌入被替换为卷积嵌入，获得局部空间信息增强，并由此去除了原始模型中的位置编码，使模型更轻量。在两个大规模权威数据集NTU-RGB+D和Kinetics-Skeleton上进行实验验证，该模型分别达到了88.1%和36.6%的Top-1精度。实验结果表明，该模型的性能有了很大的提高。

关键词：动作识别；图卷积网络；自注意力机制；Transformer

中图分类号：TN911.73; TP391.4 **文献标识码**：A **文章编号**：1009-5896(2023)04-1485-09

DOI: 10.11999/JEIT220270

Collaborative Convolutional Transformer Network Based on Skeleton Action Recognition

SHI Yuexiang ZHU Maoqing

(*School of Computer Science and Cyberspace Security*, *Xiangtan University*, *Xiangtan* 411105, *China*)

Abstract: In recent years, skeleton-based human action recognition has attracted widespread attention because of the robustness and generalization ability of skeleton data. Among them, the graph convolutional network that models the human skeleton into a spatiotemporal graph has achieved remarkable performance. However, graph convolutions learn mainly long-term interactive connections through a series of 3D convolutions, which are localized and limited by the size of convolution kernels, which can not effectively capture long-range dependencies. In this paper, a Collaborative Convolutional Transformer (Co-ConvT) network is proposed to establish remote dependencies by introducing Transformer's self-attention mechanism and combining it with Graph Convolutional Neural Networks (GCNs) for action recognition, enabling the model to extract local information through graph convolution while capturing the rich remote dependencies through Transformer. In addition, Transformer's self-attention mechanism is calculated at the pixel level, a huge computational cost is generated. The model divides the entire network into two stages. The first stage uses pure convolution to extract shallow spatial features, and the second stage uses the proposed ConvT block to capture high-level semantic information, reducing the computational complexity. Moreover, the linear embeddings in the original Transformer are replaced with convolutional embeddings to obtain local spatial information enhancement, and thus removing the positional encoding in the original model, making the model lighter. Experimentally validated on two large-scale authoritative datasets NTU-RGB+D and Kinetics-Skeleton, the model achieves respectively Top-1 accuracy of 88.1% and 36.6%. The experimental results demonstrate that the performance of

收稿日期：2022-03-14；改回日期：2022-07-07；网络出版：2022-07-21

*通信作者： 朱茂清 201921002020@smail.xtu.edu.cn

基金项目：国家自然科学基金(62172349, 62172350)，湖南省学位和研究生教育改革研究一般项目(2021JGYB085)

Foundation Items: The National Natural Science Foundation of China (62172349, 62172350), Hunan Province Degree and Postgraduate Education Reform Research General Project (2021JGYB085)

the model is greatly improved.

Key words: Action recognition; Graph Convolutional Neural Networks (GCNs); Self-attention mechanism; Transformer

1 引言

近年来，人体动作识别因其在视频监控和人机交互等领域[1]的高度实用性而受到广泛关注。基于骨骼数据的动作识别方法由于其对环境信息的鲁棒性和低成本等优点，已成为该领域最重要的研究方向之一。基于深度学习的传统方法手动将骨架构建为伪图像，并将其发送到卷积神经网络(Convolutional Neural Network, CNN)或循环神经网络(Recurrent Neural Network, RNN)进行特征提取以获得预测。然而，将骨架数据表示为2维网格并不能完全表达相关关节之间的相关性。作为以关节为顶点、骨骼为边的自然拓扑结构图，用2维图像代替图结构进行特征提取的方法无疑破坏了原有的信息相关性。因此，近年来，基于骨架的动作识别最广泛的方法已成为图神经网络，尤其是图卷积神经网络(Graph Convolutional Neural Networks, GCNs)[2]。

Yan等人[3]首先使用GCNs对人体骨骼数据进行建模，提出了ST-GCN模型，在人体关节的拓扑结构上构建空间图，并在连续帧中连接每个关节的不同位置以获得时间信息，同时聚合时空信息进行动作识别。虽然在骨架数据上表现不错，但ST-GCN仍存在着一些设计缺陷[4,5]。(1)ST-GCN仅考虑相邻范围内关节之间的联系，而对结构上距离较远但具有协同作用的关节缺乏关注。比如打篮球时，需要手、脚、腰的配合才能完成一个完整的投篮动作，而这些关节的物理距离是较远的。(2)表示人体骨骼的拓扑特征图对于所有层和动作都是固定的，这可能会影响不同网络层之间语义的丰富表示，比如网络训练后期的数据往往拥有初期所不具备的高级语义信息。(3)尽管GCN可以通过一系列3D卷积的叠加来学习长期交互联系，但这种联系是片面的、局部的，并且受到卷积核大小的限制。

最近，Transformer的成功提出了一种通过强大的自注意力机制对远程依赖进行建模的新范式[6]，虽然它最初是为自然语言处理(Natural Language Processing, NLP)任务而设计的，但人体骨骼序列的序列性和层次结构，以及Transformer在建模长期依赖方面的灵活性，使其成为解决ST-GCN弱点的完美方案，最近一些学者在图像视觉领域的研究[7–10]也证明了使用Transformer同时建模空间和时间关系的可行性，但随之而来的是需要大量的计算资源和数据才能建模长期依赖关系。

尽管Transformer在视觉任务上取得了很大成功，但在小数据集上进行训练时，其性能仍低于类似大小的CNN模型。一个可能的原因是Transformer缺乏CNN固有的一些理想特性，例如平移不变性和失真不变性。此外，CNN能够使用局部感受野、共享权重和空间子采样来捕获不同复杂度的局部空间上下文，而Transformer不具备。因此，本文提出了一个协同GCNs和Transformer的模型，在保持高效的计算和内存效率的基础上，对人类行为在空间和时间上的交互信息建模以进行动作识别。

本文将整个模型分为两个阶段：低层阶段使用纯卷积来充分学习局部空间信息，高层阶段引入Transformer来捕获远程依赖，获得全局视图以及丰富的语义信息。另外，由于Transformer天然缺少位置信息，需要使用位置嵌入来添加位置信息。针对这种情况，本文使用卷积嵌入而不是线性嵌入来学习人体序列之间的位置关系，避免使用位置编码来达到降低参数的目的。同时，这种机制使模型能够进一步捕捉局部空间上下文，减少注意力机制中的语义歧义。

本文的主要贡献总结如下：

(1) 为基于骨架的动作识别任务提出了一种协同GCNs和Transformer的模型，并将其分别应用于时间流和空间流。

(2) 设计了卷积嵌入代替原始的线性嵌入来学习位置信息，避免使用位置编码，减少模型的计算损失，大大减少了参数。

(3) 在基于骨架的动作识别的两个大规模权威数据集NTU-60和Kinetics-400上，本文的模型优于ST-GCN基线和几种最先进的方法。

2 相关工作

2.1 基于骨架的动作识别

骨架数据广泛用于动作识别，早期基于骨架的动作识别研究通常设计手工特征来建模人体[11]。然而，这些基于手工特征的方法的性能不能令人满意，因为它不能同时考虑所有因素。深度学习的发展提出了可以增强鲁棒性并获得前所未有的性能的方法，其中最广泛使用的模型是RNN和CNN。基于RNN的方法将人体关节序列建模为时间序列[12]进行计算。基于CNN的方法通过手动设计转换规则[13]将骨架数据建模为伪图像以充分利用空间信息。最近，由于人体关节与图结构的自然契合，基于GCNs的方法引起了很多关注[14–17]。

Yan等人[3]直接将骨架数据建模为图结构，每个时空图卷积层用图卷积算子构造空间特征，用卷积算子对时间动态进行建模，从空间和时间上提取特征，从而实现超越之前方法的性能。Li等人[18]在ST-GCN的基础上通过引入一种结构连接来学习一些动作相互依赖的关节之间的关系，一定程度上解决了其缺陷。

2.2 时空图卷积网络

时空图卷积网络(ST-GCN)由一系列ST-GCN块堆叠而成。每个块依次包含一个空间图卷积和一个时间卷积，用于交替提取空间和时间特征。最后一个块连接到一个全连接的Softmax分类器以生成最终预测，GCNs为每个关节引入了相邻特征的加权平均值。

令$\boldsymbol{X}_{\text{in}} \in \boldsymbol{R}^{n\times d_n}$为一帧中所有关节的输入特征，其中$d_{\text{in}}$为输入特征维度，$\boldsymbol{X}_{\text{out}} \in \boldsymbol{R}^{n\times d_{\text{out}}}$是GCNs得到的输出特征，其中$d_{\text{out}}$是输出特征维度。以上可以总结为

$$\boldsymbol{X}_{\text{out}} = \sum_{k}^{K_{\text{s}}} (\boldsymbol{X}_{\text{in}}\boldsymbol{A}_k)\boldsymbol{W}_k \tag{1}$$

$$\boldsymbol{A}_k = \boldsymbol{D}_k{}^{-\frac{1}{2}}(\tilde{\boldsymbol{A}}_k + \boldsymbol{I})\boldsymbol{D}_k{}^{-\frac{1}{2}} \tag{2}$$

其中，K_{s}是空间维度上的核大小，$\tilde{\boldsymbol{A}}_k$是表示人体关节连接的邻接矩阵，$\boldsymbol{W}_k$是可训练的权重矩阵，$\boldsymbol{I}$是单位矩阵。

时间图卷积网络是具有$(1, K_{\text{t}})$卷积核大小的标准2维卷积，最近，Shi等人[19]提出了一种自适应图结构来替代ST-GCN中传统的预定义固定图结构，使模型更加灵活，其式(1)改为

$$\boldsymbol{X}_{\text{out}} = \sum_{k}^{K_{\text{s}}} \boldsymbol{X}_{\text{in}}(\boldsymbol{A}_k + \boldsymbol{B}_k + \boldsymbol{C}_k)\boldsymbol{W}_k \tag{3}$$

其中，$\boldsymbol{A}_k$与式(1)中的相同，$\boldsymbol{B}_k$是一个可学习的数据驱动矩阵，$\boldsymbol{C}_k$是计算两个顶点相似度的矩阵。

2.3 Transformer和自注意力机制

Vaswani等人[6]提出的Transformer已成为NLP领域的主导模型，因为它在处理非常长的序列和句子的并行化方面取得了出色的效果，这是LSTM(Long and Short Term Neural Network)和RNN所不具备的。

自注意力，有时称为内部注意力，是一种将单个序列的不同位置联系起来以计算序列表示的注意力机制。自注意力层是Transformer编码器-解码器架构的构建块，其内部计算过程是：首先，对于输入数据，通过一个可训练的线性嵌入计算得到一个query向量$\boldsymbol{q} \in \boldsymbol{R}^{d_q}$，一个key向量$\boldsymbol{k} \in \boldsymbol{R}^{d_k}$和一个value向量$\boldsymbol{v} \in \boldsymbol{R}^{d_v}$，然后对$\boldsymbol{q}$向量和$\boldsymbol{k}$向量进行矩阵点乘得到相似度得分，将得分经过Softmax处理后，使用$\boldsymbol{v}$向量对其进行加权，得到最终的输出。

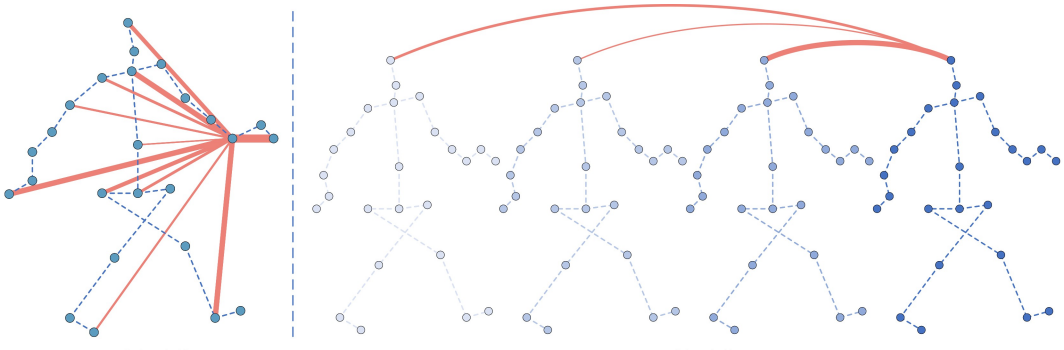
整个过程可以表示成矩阵形式为

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

其中，$\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$都是矩阵，分别表示所有打包在一起的query, key和value向量，d_k是key矩阵的维数，除以$\sqrt{d_k}$是为了防止梯度爆炸。

最近，一些研究人员尝试只使用多头自注意力层将Transformer引入计算机视觉领域[20, 21]，并在图像分类基准(例如ImageNet)上产生了最先进的结果。其他方法包括[22–24]在图像像素级别应用Transformer模型，这会产生很大的计算成本，并且必须对图像进行下采样或使用局部注意力而不是全局注意力。

在这项工作中，本文应用自注意力机制在空间和时间上建立远程连接，如图1所示。左图表示帧内连接，可以理解为捕捉人体完成动作时每个关节的协作关系；右图为各帧之间的连接，用来捕捉随着时间推移的时间顺序特征，例如，正向序列和反向序列在时间框架上的表现应该是不同的。图中连



(a) 空间连接　　(b) 时间连接

图 1 在空间流与时间流上的关节连接示意图

接线的粗细代表着不同程度的依赖，线越粗代表着两个关节之间的联系越紧密。

此外，针对计算成本的问题，本文研究了如何将GCNs和Transformer更好地结合起来，以一种轻量且有效的方式建立同时具有局部和全局感受野的模型来进行动作识别。

3 协作卷积Transformer网络

如图2所示，协作卷积Transformer网络(Collaborative Convolutional Transformer Network, Co-ConvT)是一些基本块的堆栈，它分为两个阶段，浅层阶段由自适应图卷积层组成，深层阶段由本文提出的卷积Transformer块组成，它将自注意力机制与图卷积相结合，并在空间和时间维度上交替运行以提取时空融合特征进行动作识别。

其中，L1～L5是自适应图卷积层，L6-L9是本文提出的卷积Transformer块，Pool层代表平均池化操作。

3.1 卷积Transformer块

如图3所示，一个基本块由一个空间卷积Transformer层和一个时间卷积Transformer层组成，它们分别用于提取空间和时间特征。

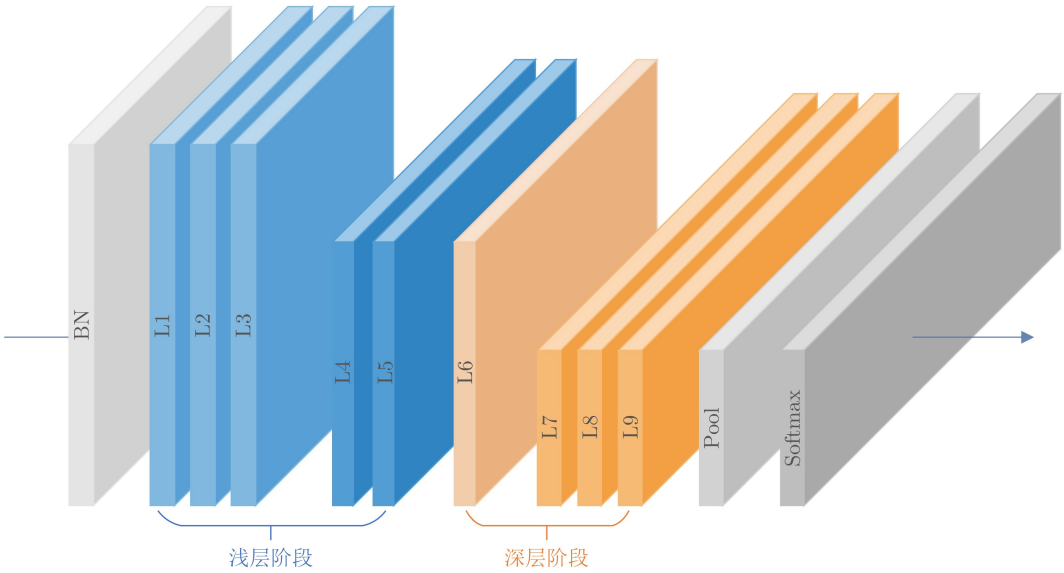


图 2 Co-ConvT网络层示意图

空间卷积Transformer层主要是让模型提取一些完成人体运动的远程协作关节的特征，重点关注在当前动作识别中发挥突出作用的关节。时间卷积Transformer层起到捕捉某些非连续帧之间联系的作用，使模型在整个动作过程中更加关注关键帧，而忽略对动作识别没有贡献的时间帧。两个卷积Transformer层之后是批归一化(Batch Normalization, BN)层和ReLU(Linear Rectification fUnction)层。此外，每个模块都增加了一个残差连接来稳定训练。

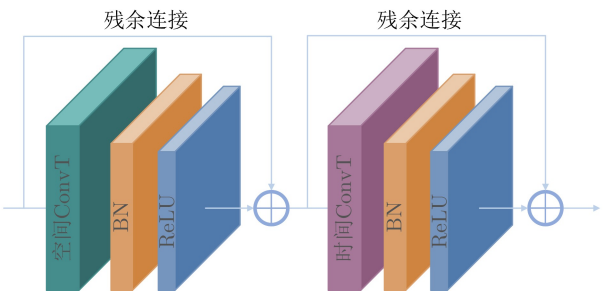


图 3 卷积Transformer基本块结构图

3.2 协作卷积Transformer层

在建模长期依赖方面，Transformer优于卷积。但是，由于Transformer的自注意力机制是在图像像素级别进行计算的，因此具有很大的计算成本。为了在两者之间取得平衡，本文提出了如图4所示的协作卷积Transformer层(ConvT)。ConvT层的输入分别送入两个支线进行处理，其上部的Transformer模块通过时空自注意力来捕捉远程依赖，获取全局视野；下部的自适应卷积模块[19]通过添加一个与原骨架等同大小的参数矩阵来跟随网络更新迭代，在网络低层提供了有助于动作识别的特征信息，学习了丰富的空间局部特征。两条支线都添加了残余连接来减少过拟合，防止梯度消失。ConvT层的后部将Transformer模块捕获的全局特征与卷积模块提取的局部特征进行矩阵加法融合，通过BN层和ReLU层激活后得到输出。

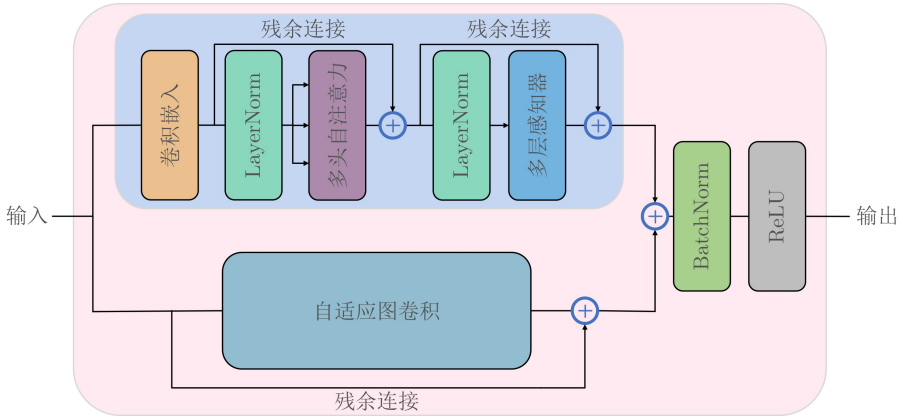


图 4 ConvT层内部框架图

3.3 Transformer模块

在模块设计上，本文沿用了原有的Transformer模型，目的是方便NLP中原始Transformer架构的迁移，使其能够高效的实现。

值得一提的是，为了使模型专注于提取空间域上的信息，本文对4维输入张量$\boldsymbol{X}_{\text{in}} \in \boldsymbol{R}^{N\times C\times T\times V}$进行降维处理得到3维张量$\boldsymbol{X}_{\text{in}} \in \boldsymbol{R}^{(N\times T)\times C\times V}$用于self-attention计算，同时，在时域上也进行了对应的处理。相比2D数据，1维序列数据更符合NLP中Transformer处理词向量的初衷，而3维张量可视为特殊的1维序列数据，由此可以更直观地拟合模型，方便后续操作。

3.3.1 卷积嵌入

Transformer模块的输入是$\boldsymbol{X} \in \boldsymbol{R}^{N\times 3\times T\times V}$，$N$表示样本数，每个样本都从原始视频中采样$T$帧，每帧包含$X$, Y, Z坐标系上的V个身体关节的坐标。在嵌入层中，本文使用具有3个卷积核的卷积操作来代替原来的线性嵌入。

具体来说，3个卷积核的大小为3×3，步长为2，padding为1，特征维度分别为$(\boldsymbol{C}_{\text{in}}, 0.5\times\boldsymbol{C}_{\text{out}})$, $(0.5\times\boldsymbol{C}_{\text{out}}, 0.5\times\boldsymbol{C}_{\text{out}})$, $(0.5\times\boldsymbol{C}_{\text{out}}, \boldsymbol{C}_{\text{out}})$，这样设计的目的是降低特征分辨率，扩大特征维度，增加其表达丰富性，使得模型可以在逐渐增大的空间足迹上表示越来越复杂的视觉模式。

此外，卷积局部感受野和zero-padding操作可以隐式地捕获位置信息，因此无需在输入中添加额外的位置编码，降低了模型复杂度，并且隐式位置信息比原始的显式位置编码更具解释性。

3.3.2 多头自注意力

多头自注意力机制意味着在h个头上并行计算，形成多个子空间来关注信息的不同方面。可以类比CNN中同时使用多个滤波器，帮助网络捕获更丰富的特征，综合利用各个方面的信息。形式上，对于输入$\boldsymbol{X}_{\text{in}}$，使用$h$个query/key/value权重矩阵将其投影到不同的表示子空间中。

$$\boldsymbol{q}_i = \boldsymbol{W}_{q,i}\text{LN}(\boldsymbol{X}_{\text{in}}) \tag{5}$$

$$\boldsymbol{k}_i = \boldsymbol{W}_{k,i}\text{LN}(\boldsymbol{X}_{\text{in}}) \tag{6}$$

$$\boldsymbol{v}_i = \boldsymbol{W}_{v,i}\text{LN}(\boldsymbol{X}_{\text{in}}) \tag{7}$$

其中，$\text{LN}(\cdot)$表示层归一化[25]，$i \in [1, 2, \cdots, h]$。然后，每个头并行应用缩放的点积注意力，最终的多头自注意力(Multi-head Self-Attention, MSA)输出是h个注意力头的串联。

$$\boldsymbol{z}_i = \text{Softmax}\left(\frac{\boldsymbol{q}_i\boldsymbol{k}_i^{\text{T}}}{\sqrt{d_k/h}}\right)\boldsymbol{v}_i, i \in [1, 2, \cdots, h] \tag{8}$$

$$\text{MSA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = \text{Concat}(\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_h)\boldsymbol{W}_{\text{out}} \tag{9}$$

4 实验

4.1 数据集

4.1.1 NTU-RGB+D 60

NTU-RGB+D数据集[26]是目前规模最大、应用最广泛的室内捕捉动作识别数据集之一。它包含从RGB+D视频样本中采样的56 880个RGB视频、深度序列、骨架数据和红外帧。骨骼信息由25个身体关节的3维坐标组成，代表60种不同的动作类别。

NTU-60数据集遵循两个不同的评估基准。Cross-Sub(X-Sub)：此基准测试中的数据集分为训练集(40 320个视频)和验证集(16 560个视频)，两个子集中的演员不同；Cross-View(X-View)：它由37 920个训练样本和18 960个测试样本组成，根据采取行动的相机不同进行拆分。

4.1.2 Kinetics-skeleton

Kinetics-skeleton数据集是从Kinetics-400[27]数据集的视频中使用OpenPose工具箱提取骨骼注释获得的。它由240 436个训练样本和19 796个测试

样本组成，共代表400个动作类。每个骨架由18个关节组成，每个关节提供2D坐标和置信度分数，对于每一帧，根据最高置信度得分最多选择2个人。

本文在训练集中训练模型，并在测试集中报告Top-1和Top-5精确度，以与之前的模型进行比较。

4.2 实验设置

本文在2块GTX-2080Ti GPU上使用PyTorch深度学习框架进行所有实验。该模型使用随机梯度下降(Stochastic Gradient Descent, SGD)和Nesterov动量(0.9)作为优化策略，使用交叉熵函数作为损失函数对梯度进行反向传播，权重衰减参数设置为0.000 1。

对于NTU数据集，总共训练了50次epoch模型，批次大小为50，初始学习率设置为0.1，并在第30次epoch和第40次epoch时减小10倍；而在Kinetics-skeleton数据集上，总共训练了65次epoch模型，批次大小为64，初始学习率为0.1，在第45次epoch和第55次epoch时减小10倍。

本文使用与2s-AGCN相同的数据预处理操作，并在前5个epoch中使用warm-up操作来预热模型。所有实验的多头注意力的头数量设置为8，Transformer的d_q, d_k和d_v的嵌入维度设置为$0.25 \times C_{\text{out}}$。

4.3 与先进模型比较

本文将模型与在Kinetics-skeleton和NTU-60数据集上的最先进方法进行比较，结果分别如表1和表2所示。在Kinetics-skeleton数据集上，作为本文实验的基线，ST-GCN的原始性能为30.7%，本文的方法将其提高到36.6%，准确率提高了5.9%。

表 1 在Kinetics-skeleton数据集上与其他模型的性能对比(%)

模型	骨骼流	Top-1精度	Top-5精度
ST-GCN[3]		30.7	52.8
AS-GCN[18]		34.8	56.5
2s-AGCN[19]	√	36.1	58.7
SAN[28]		35.1	55.7
Co-ConvT	√	**36.6**	**60.0**

表 2 在NTU-60数据集上与其他模型的性能对比(%)

模型	X-Sub基准精度	X-View基准精度
ST-GCN[3]	81.5	88.3
DPRL[29]	83.5	89.8
HCN[30]	86.5	91.1
SAN[28]	87.2	92.7
AS-GCN[18]	86.8	94.2
STA-GCN[17]	87.7	95.0
1s-Shift-GCN[4]	87.8	**95.1**
Co-ConvT	**88.1**	94.3

此外，本文采用与2s-AGCN相同的骨骼流特征来计算融合分数，结果如图5所示，无论是关节流、骨骼流还是融合流，本文的方法的性能都优于2s-AGCN模型。

另外，正如3.2节所提到的，本文模型并没有完全通过全局注意力提取特征，而是仅仅在深层网络中使用它，以此来减少参数，降低计算成本。同时，在ConvT层内部也使用了减少参数的操作，因此本文的模型以非常小的计算成本实现了良好的性能，如表3所示。

4.4 消融实验

为了验证Co-ConvT每个组件的贡献以及某些参数设置对性能的影响，本文对Kinetics-skeleton数据集进行了大量的消融实验。

首先，本文证明通过将Transformer中的线性嵌入改为卷积嵌入，可以有效增强局部空间特征的提取，并且可以隐式获取位置信息，从而去除位置编码。其次，为了平衡计算成本和模型性能，经过大量的消融实验，本文最终选择了当前的ConvT层配置，在保证计算精度的基础上减少模型参数。

4.4.1 卷积嵌入

首先，本文通过选择每个ConvT层使用卷积嵌入还是原始线性嵌入来研究提出的卷积嵌入如何影响性能，结果如表4所示。

可以观察到用卷积嵌入替换原始线性嵌入将Kinetics-skeleton上的Top-1准确率从35.2%提高到35.4%(+0.2%)，证明这种方法是一种有效的策略。然后，正如在3.3.1节提到的，由于卷积嵌入的局部感受野和zero-padding操作以一种隐式方式捕获位置信息，本文研究模型是否还需要添加位置编码。

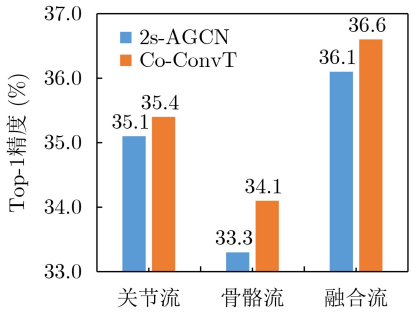


图 5 与2s-AGCN模型精度比较

表 3 在参数和精度方面与基线模型的对比

模型	参数量(10^5)	Top-1精度(%)	Top-5精度(%)
ST-GCN[3]	31.1	30.7	52.8
2s-AGCN[19]	35.5	36.1	58.7
Co-ConvT	**28.7**	**36.6**	**60.0**

通过表4中第2行和第3行的对比分析，可以证明去除位置编码并没有降低模型的性能，模型的复杂度也得到了一定程度的降低。

4.4.2 ConvT层数

如3.2节所述，本文的模型分为两个阶段。第1阶段只使用卷积操作提取特征，第2阶段使用卷积和Transformer协同提取特征。

因此，本文对如何划分这两个阶段进行了消融实验，并将网络第2阶段的ConvT层数设置为不同数量来寻求最佳的划分方式，表5显示了不同ConvT层数下的分类准确率。

可以观察到随着ConvT层数的增加，识别准确率先升高后降低。即在刚添加Transformer支线后使网络具备新层次的特征提取方式，补充了图卷积在远程依赖方面的短视，提升了性能。但随着网络层数的加深，在对实验结果中的精度和损失值进行分析后发现，训练后期网络的精度在±0.1%浮动，相应的损失值在±0.01浮动，表明此时网络已完成了对特征的学习，注意力特征图逐渐变得相似甚至几乎相同，因此一味增加层数的方式使得网络发生了“过拟合”现象，精度反而发生了下降。所以本文最终将模型的ConvT层数设置为4，在保证精度的基础上降低模型参数。

5　结束语

本文提出了一种协作卷积Transformer网络，针对图卷积在提取特征方面的片面和局部性，引入了Transformer架构中的自注意力机制来提供全局感受野，促进模型在建模全局信息方面的提升，通过协同融合Transformer捕获远程依赖和图卷积学习局部空间信息的优势来提高模型的学习能力。此外，本文在Transformer中使用卷积嵌入代替原始的线性嵌入来增强局部空间特征提取和学习空间位置信息，从而去除掉了位置嵌入，这种通过网络学习到的位置信息比Transformer中的固定位置编码更具有解释性，可以获得更丰富的语义信息，同时也降低了模型参数。

通过在NTU-RGB+D和Kinetics-skeleton数据集上进行的大量实验，证明了本文提出的Co-ConvT模型的有效性，并通过与多个主流模型进行精度对比验证了其先进性。其中，对于NTU-RGB+D数据集，Co-ConvT在X-Sub和X-View基准上的准确率分别为88.1%和94.3%，较基线模型ST-GCN分别提高了6.6%和6.0%；对于Kinetics-skeleton数据集，ST-GCN的Top-1和Top-5精度分别为30.7%和52.8%，Co-ConvT将其分别提高至36.6%和60.0%，准确率提高了5.9%和7.2%。此外，Co-ConvT的网络阶段划分策略使得网络参数量较ST-GCN减少了约2.4×10^5，较2s-AGCN减少了约6.8×10^5，一定程度上缓解了Transformer在计算代价上的压力，使得模型更为轻量化。

表 4　不同嵌入方法和移除位置编码在Kinetics-skeleton数据集上对性能的影响(%)

嵌入方法	位置编码	Top-1精度	Top-5精度
线性嵌入	×	35.2	58.1
卷积嵌入	√	35.1	57.8
卷积嵌入	×	**35.4**	**58.3**

表 5　不同ConvT层数在Kinetics-skeleton数据集的识别精度(%)

层数	Top-1	Top-5
2	35.2	57.7
3	35.5	58.1
4	35.6	58.3
5	35.4	58.0
6	35.1	57.7

参 考 文 献

[1] 石跃祥, 周玥. 基于阶梯型特征空间分割与局部注意力机制的行人重识别[J]. 电子与信息学报, 2022, 44(1): 195–202. doi: 10.11999/JEIT201006.
SHI Yuexiang and ZHOU Yue. Person re-identification based on stepped feature space segmentation and local attention mechanism[J]. *Journal of Electronics & Information Technology*, 2022, 44(1): 195–202. doi: 10.11999/JEIT201006.

[2] NIEPERT M, AHMED M, and KUTZKOV K. Learning convolutional neural networks for graphs[C]. The 33rd International Conference on International Conference on Machine Learning, New York, USA, 2016: 2014–2023.

[3] YAN Sijie, XIONG Yuanjun, and LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. The Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, USA, 2018: 912.

[4] CHENG Ke, ZHANG Yifan, HE Xiangyu, *et al*. Skeleton-based action recognition with shift graph convolutional network[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 180–189. doi: 10.1109/CVPR42600.2020.00026.

[5] LIU Ziyu, ZHANG Hongwen, CHEN Zhenghao, *et al*. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 140–149. doi: 10.1109/CVPR42600.2020.00022.

[6] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[7] MEINHARDT T, KIRILLOV A, LEAL-TAIXE L, *et al*. TrackFormer: Multi-object tracking with transformers[J]. arXiv: 2101.02702, 2021. doi: 10.48550/arXiv.2101.02702..

[8] SUN Peize, CAO Jinkun, JIANG Yi, *et al*. TransTrack: Multiple object tracking with transformer[J]. arXiv: 2012.15460, 2020. doi: 10.48550/arXiv.2012.15460..

[9] ZHENG CE, ZHU Sijie, MENDIETA M, *et al*. 3D human pose estimation with spatial and temporal transformers[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 11636–11645. doi: 10.1109/ICCV48922.2021.01145.

[10] CHU Peng, WANG Jiang, YOU Quanzeng, *et al*. TransMOT: Spatial-temporal graph transformer for multiple object tracking[J]. arXiv: 2104.00194, 2021. doi: 10.48550/arXiv.2104.00194..

[11] FERNANDO B, GAVVES E, ORAMAS J M, *et al*. Modeling video evolution for action recognition[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 5378–5387. doi: 10.1109/CVPR.2015.7299176.

[12] LI Shuai, LI Wanqing, COOK C, *et al*. Independently Recurrent Neural Network (IndRNN): Building a longer and deeper RNN[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5457–5466. doi: 10.1109/CVPR.2018.00572.

[13] LI Chao, ZHONG Qiaoyong, XIE Di, *et al*. Skeleton-based action recognition with convolutional neural networks[C]. 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 2017: 597–600. doi: 10.1109/ICMEW.2017.8026285.

[14] ZHANG Pengfei, LAN Cuiling, ZENG Wenjun, *et al*. Semantics-guided neural networks for efficient skeleton-based human action recognition[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 1109–1118. doi: 10.1109/CVPR42600.2020.00119.

[15] ZHANG Xikun, XU Chang, and TAO Dacheng. Context aware graph convolution for skeleton-based action recognition[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 14321–14330. doi: 10.1109/CVPR42600.2020.01434.

[16] 曾胜强, 李琳. 基于姿态校正与姿态融合的2D/3D骨架动作识别方法[J]. 计算机应用研究, 2022, 39(3): 900–905. doi: 10.19734/j.issn.1001-3695.2021.07.0286.

ZENG Shengqiang and LI Lin. 2D/3D skeleton action recognition based on posture transformation and posture fusion[J]. *Application Research of Computers*, 2022, 39(3): 900–905. doi: 10.19734/j.issn.1001-3695.2021.07.0286.

[17] 李扬志, 袁家政, 刘宏哲. 基于时空注意力图卷积网络模型的人体骨架动作识别算法[J]. 计算机应用, 2021, 41(7): 1915–1921. doi: 10.11772/j.issn.1001-9081.2020091515.

LI Yangzhi, YUAN Jiazheng, and LIU Hongzhe. Human skeleton-based action recognition algorithm based on spatiotemporal attention graph convolutional network model[J]. *Journal of Computer Applications*, 2021, 41(7): 1915–1921. doi: 10.11772/j.issn.1001-9081.2020091515.

[18] LI Maosen, CHEN Siheng, CHEN Xu, *et al*. Actional-structural graph convolutional networks for skeleton-based action recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 3590–3598. doi: 10.1109/CVPR.2019.00371.

[19] SHI Lei, ZHANG Yifan, CHENG Jian, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 12018–12027. doi: 10.1109/CVPR.2019.01230.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[C/OL]. The 9th International Conference on Learning Representations, 2021.

[21] TOUVRON H, CORD M, DOUZE M, *et al*. Training data-efficient image transformers & distillation through attention[L/OL]. The 38th International Conference on Machine Learning, 2021: 10347–10357.

[22] RAMACHANDRAN P, PARMAR N, VASWANI A, *et al*. Stand-alone self-attention in vision models[C]. *Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada,* 2019: 7.

[23] SHARIR G, NOY A, and ZELNIK-MANOR L. An image is worth 16x16 words, what is a video worth?[J]. arXiv: 2103.13915, 2021. doi: 10.48550/arXiv.2103.13915..

[24] PLIZZARI C, CANNICI M, and MATTEUCCI M. Spatial temporal transformer network for skeleton-based action

recognition[C]. International Conference on Pattern Recognition, Milano, Italy, 2021: 694–701. doi: 10.1007/978-3-030-68796-0_50.

[25] BA J L, KIROS J R, and HINTON G E. Layer normalization[J]. arXiv: 1607.06450, 2016. doi: 10.48550/arXiv.1607.06450..

[26] SHAHROUDY A, LIU Jun, NG T T, *et al*. NTU RGB+D: A large scale dataset for 3D human activity analysis[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1010–1019. doi: 10.1109/CVPR.2016.115.

[27] KAY W, CARREIRA J, SIMONYAN K, *et al*. The kinetics human action video dataset[J]. arXiv: 1705.06950, 2017. doi: 10.48550/arXiv.1705.06950..

[28] CHO S, MAQBOOL M H, LIU Fei, *et al*. Self-attention network for skeleton-based human action recognition[C]. 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass, USA, 2020: 624–633. doi: 10.1109/WACV45572.2020.9093639.

[29] TANG Yansong, TIAN Yi, LU Jiwen, *et al*. Deep progressive reinforcement learning for skeleton-based action recognition[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 5323–5332. doi: 10.1109/CVPR.2018.00558.

[30] LI Chao, ZHONG Qiaoyong, XIE Di, *et al*. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[C]. The Twenty-Seventh International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 786–792. doi: 10.24963/ijcai.2018/109.

石跃祥：男，教授，硕士生导师，研究方向为图像处理和行为识别.

朱茂清：男，硕士，研究方向为动作识别.

责任编辑：马秀强