基于多流空间注意力图卷积SRU网络的骨架动作识别

赵俊男,佘青山,孟 明,陈 云
(杭州电子科技大学自动化学院,浙江杭州 310018)

摘 要: 基于骨架的动作识别越来越受到重视.针对现有算法推理速度慢、数据模式单一等问题,本文提出了一种轻量且高效的方法.该网络在简单循环单元(Simple Recurrent Unit,SRU)中嵌入图卷积算子构建图卷积SRU(GC-SRU)模型,来捕获数据的时空域信息.同时,为了加强节点间的区分,采用空间注意力网络和多流数据融合方式,将GC-SRU拓展成多流空间注意力图卷积SRU(MSAGC-SRU).最后,在公开数据集上进行实验分析.结果表明,本文方法在Northwestern-UCLA上的分类准确率达到了93.1%,模型FLOPs为4.4 G;NTU RGB+D上的分类准确率在CV、CS评估协议下分别达到92.7%和87.3%,模型FLOPs为21.3 G,达到了计算效率和分类精度的良好平衡.

关键词: 动作识别;图卷积;注意力机制;数据融合
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2022)07-1579-07
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210416

Skeleton Action Recognition Based on Multi-Stream Spatial Attention Graph Convolutional SRU Network

ZHAO Jun-nan,SHE Qing-shan,MENG Ming,CHEN Yun
(*College of Automation,Hangzhou Dianzi University,Hangzhou,Zhejiang* 310018,*China*)

Abstract: Action recognition with skeleton data has attracted more attention. In order to solve the problems of low reasoning speed and single data mode of most algorithms, a lightweight and efficient method is proposed. The network embeds the graph convolution operator in the simple recurrent unit(SRU) to construct the graph convolutional SRU(GC-SRU), which can capture the spatial-temporal information of data. Meanwhile, to enhance the distinction between nodes, spatial attention network and multi-stream data fusion are used to expand GC-SRU into multi-stream spatial attention graph convolutional SRU(MSAGC-SRU). Finally, the proposed method is evaluated on two public datasets. Experimental results show that the classification accuracy of our method on Northwestern-UCLA reaches 93.1% and the FLOPs of the model is 4.4 G. The accuracy on NTU RGB+D reaches 92.7% and 87.3% under the CV and CS evaluation protocols, respectively, and the FLOPs of the model is 21.3 G. The proposed model has achieved good trade-off between computational efficiency and classification accuracy.

Key words: action recognition;graph convolution;attention mechanism;data fusion

1 引言

人类行为的识别是计算机视觉中一项基本又富有挑战性的任务,促进了许多应用的产生,如人机交互、异常行为检测等[1].与传统利用RGB图片或视频流进行动作识别的方法相比,基于骨架的动作识别不受背景杂波、光照变化等限制,对目标动作的表示更加健壮.

近年来,研究人员充分挖掘骨架数据的图结构信息,开始将图卷积网络推广到骨架图上,来进行动作识别任务.Yan等[2]提出了基于图的动态骨架建模通用公式,提出时空图卷积(Spatio-Temporal Graph Convolutional Networks,ST-GCN)网络,应用于骨架动作识别.ST-GCN采用空间图卷积捕捉骨骼数据的空间信息,利用时间卷积在时间维度上对数据建模,获得了更好的

收稿日期:2021-03-29;修回日期:2021-12-30;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61871427);浙江省自然科学基金重点项目(No.LZ22F010003)

识别效果. 而 Si 等[3]发挥了长短时记忆网络(Long Short Time Memory,LSTM)强大的序列建模性能,并将空间图卷积操作嵌入LSTM的门计算中,提出了注意力增强图卷积LSTM网络(AGC-LSTM),取得了比ST-GCN更好的效果. Lei 等[4]则更多地关注骨骼数据蕴含的二阶信息,包括骨骼的长度和方向,提出了多流注意增强型自适应图卷积网络(MS-AAGCN). 该方法增加了骨骼图构造模型的灵活性,增强了模型的泛化能力,显著提高了识别精度. 然而,上述算法为了达到更高的识别精度,采取叠加多层网络或者使用时间维度建模能力强的模块,会导致模型在训练和测试中更耗时. 因此,以上算法仍然存在模型推理速度慢、计算复杂度高等问题,值得进一步研究.

最近,Lei 等[5]提出了一种简单循环单元(Simple Recurrent Unit,SRU)结构,具有比LSTM更高的并行性和更快的推理速度. She 等[6]提出一种全局上下文注意力时空SRU模型(GCA-ST-SRU). 该方法首先通过SRU单元来构建模型,以克服LSTM推理速度慢等的问题. 但是该方法未考虑骨骼间的结构联系,而且对数据的利用比较单一.

为了解决上述方法的不足,本文提出了一种新的多流空间注意力图卷积SRU网络(MSAGC-SRU)方法. 首先,用图卷积算子替换SRU模型中的单元结构门的全连接计算,提出图卷积SRU(GC-SRU)方法. 同时对数据的空间和时间域进行建模. 然后,为了加强关节间的区分性,进一步引入了空间注意力网络,构建空间注意力图卷积SRU(SAGC-SRU)网络,使模型更加关注重要的关节. 最后,受MS-AAGCN方法中多流数据输入方式的启发,进一步提出一种简单高效的多流数据融合方式,该方式将多数据流一次输入、同时训练,能显著减少运行时间. 最后在两个公共数据集上的实验表明,所提方法在识别精度和计算效率间达到了良好的平衡.

2 本文算法

基于骨架的动作识别是对骨架序列进行建模,并学习序列的时间动力学信息. 本文提出一种端到端的多流空间注意力图卷积SRU网络(MSAGC-SRU),用于骨架动作识别任务,其网络架构如图1所示.

对于骨架序列,本文首先通过多流数据融合方法,得到多流融合数据. 然后,设计三层SAGC-SRU网络来模拟时空特征,如图2所示. 其中,图卷积算子能够捕捉骨骼数据的空间结构信息和时间动态的判别特征. 同时,在SAGC-SRU层的输出位置引入空间注意力网络,增加了对重要关节点的关注度. 参考Si等人[3]的设置,本文在SAGC-SRU数据输入时,在时间维度上进行

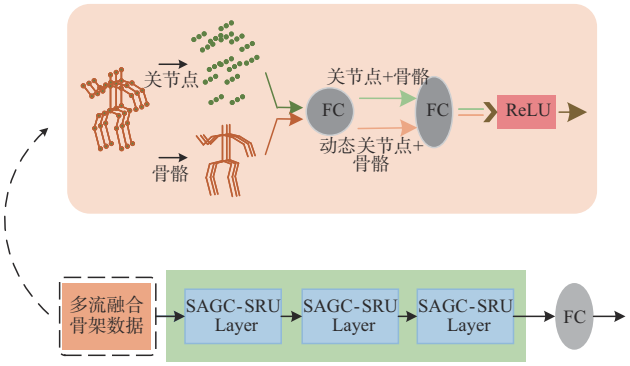


图1 MSAGC-SRU网络架构

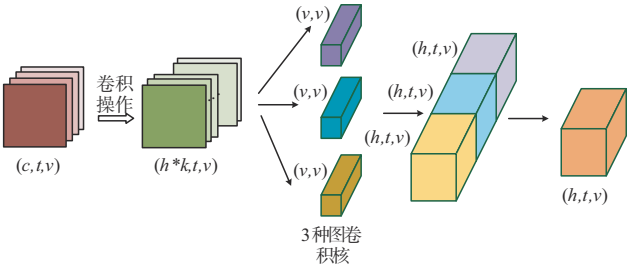


图2 GCN可视化过程

平均池化操作. 能够在时间域上压缩输入数据和参数的量,增加SAGC-SRU顶层的时间感受野,同时减小网络过拟合的可能性,并加快了模型的推理速度. 最后是全连接层,实现动作的分类任务.

2.1 多流数据融合

骨架数据的一阶(关节点坐标)和二阶信息(骨骼的方向和长度)[4]以及它们的动态信息都是区分不同动作的重要信息. 因此,有必要对骨架数据的多种模式建模. 受到Lei等[4]相关工作的启发,本文使用4种模式的数据流,分别为以原始关节点坐标为输入的节点流,以关节点空间坐标的差分为骨骼流,以及基于节点流和骨骼流数据在时间维度上的差分得到的动态数据流.

一般定义靠近骨架重心的关节点为父关节i,远离重心的为子关节j,在第t帧的骨架中,设父关节$v_{i,t}=(x_{i,t},y_{i,t},z_{i,t})$和子关节$v_{j,t}=(x_{j,t},y_{j,t},z_{j,t})$,则骨骼可表示为

$$e_{i,j,t}=(x_{j,t}-x_{i,t},y_{j,t}-y_{i,t},z_{j,t}-z_{i,t}) \tag{1}$$

通过式(1)对原始的节点流数据处理,可以得到骨骼流数据. 设$v_{i,t}=(x_{i,t},y_{i,t},z_{i,t})$为第$t$帧的关节,$v_{i,t+1}=(x_{i,t+1},y_{i,t+1},z_{i,t+1})$为第$t+1$帧的关节,则两帧间节点的动态流信息可以表示为

$$m_{i,t,t+1}=(x_{i,t+1}-x_{i,t},y_{i,t+1}-y_{i,t},z_{i,t+1}-z_{i,t}) \tag{2}$$

本文没有采用将各流数据依次训练后,再将训练结果融合的方法. 而是在一次训练中同时输入多流融合数据,减少实验次数,减少训练时间. 如图1的多流数据融合部分所示,具体步骤如下:

步骤1:原始骨架序列数据为节点流数据,经过上

述骨骼的定义式(1),得到骨骼流数据;

步骤2:将Step1获得的两流数据拼接后,通过全连接层,将信息编码成高维向量,作为数据的融合特征;

步骤3:通过步骤Step2得到的两流融合数据,获取其动态流数据.再和它拼接后经过全连接层和ReLU激活函数,得到目标的多流融合数据.

2.2　空间注意力图卷积SRU

图卷积神经网络(Graph Convolutional Network,GCN)是一种通用而有效的图结构数据学习表示框架[3].在基于骨架的动作识别中,设$\varsigma_t=\{v_t,\varepsilon_t\}$表示第$t$帧的人体骨骼图,$v_t=\{v_{t1},v_{t2},...,v_{tN}\}$是$N$个关节点的集合,则$\varepsilon_t=\{(v_{ti},v_{tj}):v_{ti},v_{tj}\in v_t,v_{ti}\sim v_{tj}\}$可以表示成骨骼边的集合,$v_{ti}\sim v_{tj}$表示节点$i$和节点$j$是无向边连接.邻接矩阵$\boldsymbol{A}_t$可以通过$\varepsilon_t$指定:

$$\boldsymbol{A}_t(i,j)=\begin{cases}1, & (v_{ti},v_{tj})\in\varepsilon_t\\ 0, & \text{其他}\end{cases}\tag{3}$$

在一个骨架图上,定义节点v_{ti}的邻集$B(v_{ti})=\{v_{tj}|d(v_{tj},v_{ti})\le D\}$,其中$d(v_{tj},v_{ti})$表示从节点$v_{ti}$到$v_{tj}$的任何路径的最小长度.可以给出在点$v_{ti}$上的图卷积公式:

$$f_{\text{out}}(v_{ti})=\sum_{v_{tj}\in B(v_{ti})}\frac{1}{Z_{ti}(v_{tj})}\boldsymbol{X}(v_{tj})\boldsymbol{W}(\ell(v_{tj}))\tag{4}$$

其中,$\boldsymbol{X}(v_{tj})$表示节点v_{tj}的特征,$\boldsymbol{W}(\cdot)$是一个权重函数,v_{tj}为与v_{ti}距离为1的相邻节点.$\ell(\cdot)$是一个映射函数.因为邻集$B(v_{ti})$的节点数量是变化的,权重函数数量是固定的,所以需要将所有相邻的节点映射到一个固定标签的子集[4]中,每个子集都有一个唯一的关联权重向量.本文参考了ST-GCN[2]的映射策略,将$B(v_{ti})$分为三个子集:第一个子集为v_{tj}本身,第二个为空间位置上比v_{tj}更靠近骨架重心的邻点集合,第三个则为更远离重心的邻点集合,$Z_{ti}(v_{tj})$即为对应子集个数.

引入邻接矩阵,图卷积可以表示为:

$$\boldsymbol{f}_{\text{out}}=\sum_{k=1}^{k}\boldsymbol{\Lambda}_k^{-\frac{1}{2}}\boldsymbol{A}_k\boldsymbol{\Lambda}_k^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}_k\tag{5}$$

其中,$k\in\{1,2,\cdots,k\}$是根据映射策略得到的子集的固定标签.$\boldsymbol{A}_k$是标签k空间构型中的邻接矩阵,$\boldsymbol{\Lambda}_k^{ij}=\sum_j\boldsymbol{A}_k^{ij}$是一个度矩阵.为了更好地理解图卷积在骨架序列数据中的计算过程,图2给出可视化形式.图中c表示骨架数据的输入通道数,t是数据的序列长度,v是骨架数据中的关节点数量,h表示输出通道.

本文选择GCN与SRU结合.SRU是一种简单的循环单元模型[7],它解决了因反向传播引起的梯度消失问题[8],并通过在门输入中隐去前一时刻的隐藏状态实现计算上的并行性[9],具有更快的训练速度和推理速度.从Di等人[10]的思路来看,SRU需要堆叠更多层,才能获得与单层LSTM相似的性能.本文将SRU应用到骨架动作识别中,需要对SRU进行改进.

本文使用空间注意力图卷积SRU(SAGC-SRU),对骨架序列数据进行时空域建模[11],结构如图3所示.将SRU的输入门、遗忘门、重置门的全连接计算替换成图卷积计算,捕获数据的空间信息,并在隐藏状态$\hat{\boldsymbol{H}}_t$引入空间注意力机制,来关注不同重要程度的关节点.

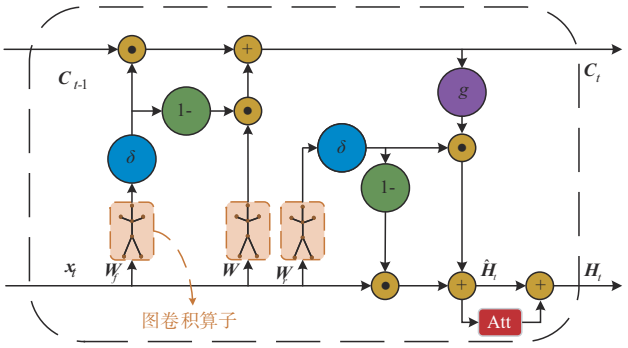


图3　SAGC-SRU模型结构

SAGC-SRU的输入$\boldsymbol{x}_t$、隐藏状态$\hat{\boldsymbol{H}}_t$和存储状态$\boldsymbol{C}_t$都是图结构数据.其功能定义如下:

$$\tilde{\boldsymbol{x}}_t=\text{AveragePooling}(\boldsymbol{x}_t)\tag{6}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f*\tilde{\boldsymbol{x}}_t+\boldsymbol{b}_t)\tag{7}$$

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_r*\tilde{\boldsymbol{x}}_t+\boldsymbol{b}_r)\tag{8}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t\odot\boldsymbol{C}_{t-1}+(\boldsymbol{1}-\boldsymbol{f}_t)\odot(\boldsymbol{W}*\tilde{\boldsymbol{x}}_t)\tag{9}$$

$$\hat{\boldsymbol{H}}_t=\boldsymbol{r}_t\odot g(\boldsymbol{C}_t)+(\boldsymbol{1}-\boldsymbol{r}_t)\odot\tilde{\boldsymbol{x}}_t\tag{10}$$

$$\boldsymbol{H}_t=f_{\text{att}}(\hat{\boldsymbol{H}}_t)+\hat{\boldsymbol{H}}_t\tag{11}$$

其中,AveragePooling表示平均池化操作,使用Pytorch中的F.avg_pool1d函数对输入$\boldsymbol{x}_t$进行平均池化.$*$表示图卷积算子,$\boldsymbol{W}*\tilde{\boldsymbol{x}}_t$表示$\boldsymbol{W}$和$\tilde{\boldsymbol{x}}_t$的图卷积.$\odot$表示哈达玛积,$\sigma(\cdot)$是Sigmoid激活函数,$g(\cdot)$表示tanh激活函数,$f_{\text{att}}(\cdot)$是选择关键节点信息的空间注意力网络.引入注意力机制后,输出$\boldsymbol{H}_t$能增强关键点的信息,且不会削弱非关键点的信息,来保持空间信息的完整性[3].

2.3　空间注意力网络

一般情况下,不同动作涉及到不同关节点,不同关节点的重要程度就会不同.由此,本文提出一种空间注意力模型(Spatial Attention Model)来区分节点的重要程度,模型结构如图4所示.该模型引入了软注意力机制[12],为帧内每个关节点分配一个空间注意权重,使得模型能够自适应地将注意力集中在某些关节上.

空间注意力网络的输入是图卷积SRU(GC-SRU)的隐藏状态$\hat{\boldsymbol{H}}_t$,它含有骨架动作序列的时态信息和丰富的空间结构信息,有利于关键关节点的选择[13].首先将每个节点的输出信息通过一层全连接层和ReLU激

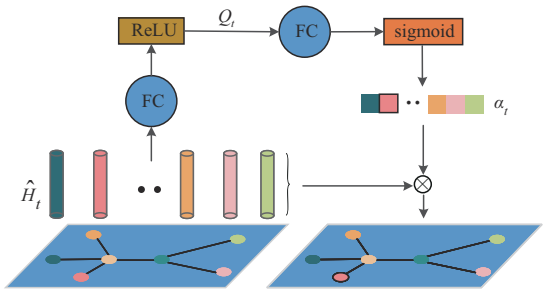


图4 空间注意力网络

活函数，聚合成一个查询向量$\boldsymbol{Q}_t$[2]：

$$\boldsymbol{Q}_t=\mathrm{ReLU}(\sum_{i=1}^{N}\boldsymbol{W}\hat{\boldsymbol{H}}_{ti}) \tag{12}$$

其中，$\alpha_t=(\alpha_{t1},\alpha_{t2},\cdots,\alpha_{tN})$表示节点的注意力分数．全连接层防止了网络复杂化也提升了模型的泛化能力，最后利用激活函数得到注意力分数．

经过最后一层SAGC-SRU后，本文将所有节点和注意力分数加权后的特征聚合得到$\boldsymbol{F}_t$，将每个时间步长的聚合特征$\boldsymbol{F}_t$转换成C类的分数$\boldsymbol{o}_t$，其中$\boldsymbol{o}_t=\{o_{t1},o_{t2},\cdots,o_{tC}\}$，$i$类的预测可以表示成式(14)：

$$\boldsymbol{F}_t=\sum_{i=1}^{N}\alpha_{ti}\cdot\hat{\boldsymbol{H}}_{ti} \tag{13}$$

$$\hat{y}_{ti}=\frac{e^{o_{ti}}}{\sum_{j=1}^{C}e^{o_{tj}}},\quad i=1,\cdots,C \tag{14}$$

本文采用下面的损失函数来监督所建模型：

$$L=-\sum_{i=1}^{C}y_i\log\hat{y}_i+\lambda\sum_{j=1}^{3}\sum_{n=1}^{N}\left(1-\frac{\sum_{t=1}^{T_j}\alpha_{tnj}}{T_j}\right)^2 +\beta\sum_{j=1}^{3}\frac{1}{T_j}\sum_{t=1}^{T_j}(\sum_{n=1}^{N}\alpha_{tnj})^2 \tag{15}$$

其中，T_j表示第j层SAGC-SRU上的时间步长，$y=\{y_1,y_2,\cdots,y_C\}$表示真实标签，与预测标签$\hat{y}$进行对比．λ和β是权重衰减系数，平衡了正则化项的作用，减轻了反向传播中的梯度消失和过拟合现象．本文分别设置λ和β为0.01和0.001．

3 实验结果与分析

3.1 数据集和评价指标

Northwestern-UCLA数据集包含1 494个样本，分别由三台Kinect摄像头同时从三个不同的视角拍摄．一共涵盖10个动作类别，分别是单手捡起、双手捡起、丢垃圾、四处走动、坐下、起立、穿、脱、扔、搬运．本实验采用了文献[14]中的评估方案：使用前两个摄像机视角的样本作为训练集，另一个视角的样本作为测试集．

NTU RGB+D数据集包含56 880个样本[12]，包含60种行为，包含了日常、相互和健康相关的行为．对于该数据集，本文采用跨受试者(Cross Subjects，CS)和跨视角(Cross Views，CV)评估协议．在CS协议中，由20名受试者采集的数据构成训练集，其余20名受试者用于测试．对于CV协议中，前两个视角捕获的样本用于训练，另一个视角的数据用作测试．

为了评估算法在分类任务的有效性，本文使用主流的评估指标：分类准确率(Accuracy)和算法的浮点运算数(FLOPs)．在Northwestern-UCLA数据集的实验中，实验增加了训练1 000次训练和测试样本的耗时统计．

根据文献[15]提供的计算方法，FLOPs值计算与数据集的动作类别C、骨节点数量N、骨架序列长度T以及网络层类型等有关．以本文提出的MSAGC-SRU模型和Northwestern-UCLA Dataset为例，其中$N=20$，$T=52$．该模型的FLOPs计算包含三部分：多流数据融合部分，三层SAGC-SRU网络层和最后的全连接分类层．

首先是多流数据融合部分，有两层的全连接层．根据实验的超参数设置，它们的FLOPs分别为$2\times3\times256\times N\times T$和$2\times C^2\times N\times T$，其中$C=512$．然后是三层SAGC-SRU．每一层的SAGC-SRU由SRU、图卷积、空间注意力机制三部分FLOPs相加得到．SRU的FLOPs可表示为$(3\times2C^2)\times N\times T$．图卷积的计算过程已通过图2可视化，其将骨骼节点按分区规则分为三部分的图卷积操作．代码实现中，使用了Pytorch中的torch.matmul函数，可视为矩阵相乘的操作．图卷积部分的FLOPs可表示为$3\times(NC^2+N^2C)\times T\times2$．空间注意力部分使用了全连接层，FLOPs可表示为$2\times2\times C^2\times N\times T$．因为在时间维度上使用了平均池化，三层网络的$T$分别为26，13，7，$C=512$．最后一层是全连接层，FLOPs为$2\times C\times10$，其中$C=512$．在计算完每部分的FLOPs后，将其相加后乘以数据集包含的人数即为模型的FLOPs．

3.2 实验细节

实验使用Northwestern-UCLA和NTU RGB+D两个公共数据集，将抽取每个骨架序列中的固定长度T，其中Northwestern-UCLA数据集的T设置为50，NTU RGB+D数据集设为100．训练使用Adam优化器来优化网络[16]，Dropout设为0.5，初始学习率分别设为0.001和0.01，weight-decay为1e−5，训练的批次大分别设为32和64．网络设置三层SAGC-SRU网络，实验均在Pytorch框架上进行，GPU为2块NVIDIA GTX 1080ti．

3.3 实验结果

3.3.1 Northwestern-UCLA Dataset

为了验证本文所提方法的有效性，首先在Northwestern-UCLA数据集上与同领域的其他先进方法进行对比．这里将要比对的方法包括：基于人工提取特

征的方法[17,18]、CNN和RNN方法[19,20]、改进型LSTM方法[3,21]. 结果如表1所示. 较于传统的手工提取特征方法和经典的RNN、CNN方法,本方法在分类精度方面优势明显. 本文提出的MSAGC-SRU方法与Ensemble TS-LSTM[21]方法相比,在分类精度上提高了3.9%,仅仅比AGC-LSTM方法低了0.2%,但是FLOPs值低于AGC-LSTM的1/2. 初步说明,MSAGC-SRU方法在大大减小计算复杂度的同时,能够达到良好的分类精度.

表1　Northwestern-UCLA数据集上实验结果比较

Methods	Accuracy(%)	FLOPs(G)
Lie group[17]	74.2	—
Actionletensemble[18]	76.0	—
HBRNN-L[19]	78.5	—
VisualizaionCNN[20]	86.1	—
Ensemble TS-LSTM[21]	89.2	—
AGC-LSTM[3]	93.3	10.9
MSAGC-SRU	93.1	4.4

为了说明本方法中采用的多流数据融合方法和空间注意力机制在实验中的有效性,再进行图卷积、数据融合方式和空间注意力机制的消融实验,以SRU作为基线方法,并与AGC-LSTM方法进行比较,实验结果如表2所示. 以下实验方法将主要用于消融实验对比:

(1) SRU:标准的3层SRU堆叠网络;

(2) GC-SRU:3层GC-SRU网络,输入单流数据;

(3) MSGC-SRU:3层的GC-SRU网络,输入多流融合数据;

(4) SAGC-SRU:3层SAGC-SRU网络,输入单流数据;

(5) MSAGC-SRU:3层的SAGC-SRU网络,输入多流融合数据.

表2给出了本文方法在Northwestern-UCLA数据集上的消融实验结果. 只加入图卷积运算的GC-SRU方法的分类准确率为84.8%,比单纯的SRU提高了3.5%,说明图卷积的应用提高了分类性能. 当GC-SRU使用多流融合数据时,其分类结果达到了90.3%,在单流输入的基础上提高了5.5%. 在此基础上加入空间注意力机制,得到的SAGC-SRU以及MSAGC-SRU,分类精度分别为90.1%和93.1%. 说明空间注意力机制同样带来比较大的性能提升. 消融实验的结果表明,MSAGC-SRU方法中使用的图卷积、多流数据融合以及空间注意力机制都能提升模型的分类性能.

下面,本文获取算法的训练和测试时间,能直观地看出MSAGC-SRU的轻量级优势. 表3给出不同方法连续训练和测试1 000次的耗时. 每次测试时只输入一个样本,并重复5次实验,取测试时间的平均值. 可以看出,SRU的训练和测试速度是最快的,使用多流融合模块、图卷积运算和注意力机制都会增加模型的训练和测试耗时. 与AGC-LSTM方法相比,提出的MSAGC-SRU方法在该条件下的训练效率提高了2.1倍,测试效率提升了2倍.

表2　Northwestern-UCLA数据集上消融实验结果

Methods	Accuracy(%)	FLOPs(G)
SRU	81.3	—
GC-SRU	84.8	3.0
MSGC-SRU	90.3	3.5
SAGC-SRU	90.1	3.9
MSAGC-SRU	93.1	4.4

表3　在Northwestern-UCLA上测试不同算法训练和测试1 000次样本的时间

Methods	Training time(s)	Testing time(s)
SRU	5.93	1.00
GC-SRU	6.08	1.86
MSGC-SRU	6.38	2.04
AGC-LSTM[3]	14.30	4.74
SAGC-SRU	6.49	2.04
MSAGC-SRU	6.77	2.36

表1~表3的结果表明,MSAGC-SRU方法在Northwestern-UCLA数据集上,不仅在动作分类任务中表现出色,还大幅提高了模型的训练和测试效率,在分类准确率和计算成本之间有很好的平衡.

3.3.2　NTU RGB+D Dataset

为了验证本文方法的泛化能力,继续在更大的数据集NTU RGB+D上实验. 比较的方法包括ST-GCN[2]、2s-Adaptiva GCN[22]、AGC-LSTM[3]和MS-AAGCN[4],实验结果如表4所示.

表4　在NTU RGB+D数据集上实验结果比较

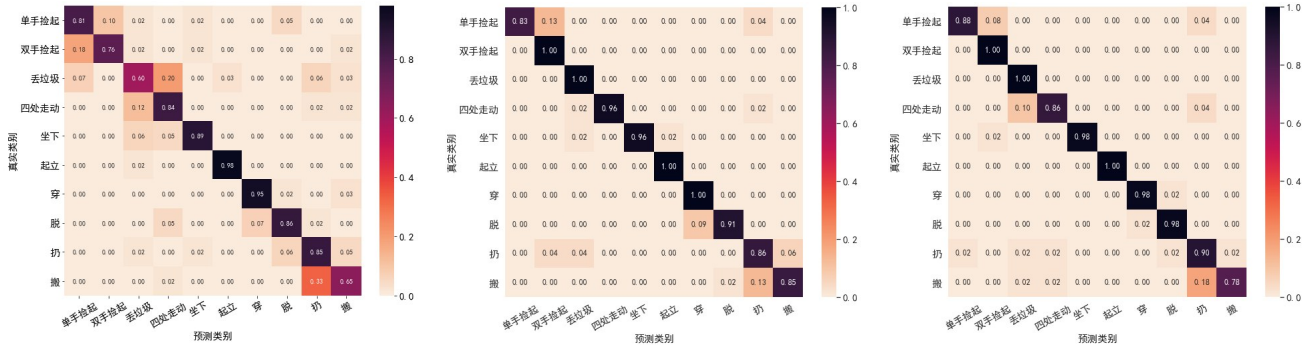
Methods	CV(%)	CS(%)	FLOPs(G)
HBRNN-L[19]	64.0	59.1	—
Deep-LSTM[23]	67.3	60.7	—
ST-GCN[2]	88.3	81.5	16.2
AGC-LSTM[3]	95.0	89.2	54.4
2s-Adaptive GCN[22]	95.1	88.5	35.8
MS-AAGCN[4]	96.2	90.0	71.6
MSAGC-SRU	92.7	87.3	21.3

从表4中可以看出,在CV协议中,本文提出的MSAGC-SRU方法比ST-GCN提高了4.4%的分类准确率;在CS协议中,本文方法提高了5.8%的准确率. 与AGC-LSTM[3]相比,CV协议中MSAGC-SRU的分类精度低了2.3%,CS协议中低了1.9%,但FLOPs低于AGC-LSTM的1/2. 与更先进的算法MS-AAGCN相比,CV协议下低了3.5%,CS协议中低了2.7%,但是从算法复杂

度看，本文方法的FLOPs仅为MS-AAGCN的1/3. 结果表明，MSAGC-SRU方法在类别增多、数据特征增加的情况下分类准确率有明显损失，但是显著降低了算法复杂度. 在重视设备的运行效率的场景下，比如嵌入式或移动端的应用中，MSAGC-SRU能尽可能保持损失很小的精度来满足非常高效的应用任务.

4 讨论

在Northwestern-UCLA数据集的实验中，本文用混淆矩阵分析了训练模型的测试分类结果. 从图5中可以看到，SRU方法在“单手捡起”和“双手捡起”两种动作间分类精度不高，在“丢垃圾”和“四处走动”两个动作也不能很好的区分，尤其是“扔”的动作识别效果很差，主要因为这些动作间的相似度较高，普通的SRU方法不能很好的区分. 而本文提出的MSAGC-SRU方法很好的提高了对这些动作的分类精度，即使与AGC-LSTM相比，在“自处走动”、“搬”等动作的分类精度会略微高一些. 这些结果表明，MSAGC-SRU方法在骨架动作识别任务中是一种有效的方法.



(a) SRU方法 (b) MSAGC-SRU方法 (c) AGC-LSTM方法

图5 Northwestern-UCLA数据集上测试结果的混淆矩阵对比图

5 结论

本文提出了一种多流空间注意力图卷积SRU(MSAGC-SRU)方法，并应用于基于骨架的三维动作识别. MSAGC-SRU方法将图卷积算子代替门的全连接运算，利用数据的空间信息. 同时，引入空间注意力网络，增强节点的关注度. 此外，为了表现骨骼数据的多样性，采用多流融合数据输入. 在两个公开数据集上的实验表明，与先进方法比，所提出的MSAGC-SRU提高了数倍的计算效率，同时保持了不错的分类准确率. 接下来研究的开展会将该算法应用在嵌入式平台上，开发相关应用.

参考文献

[1] 罗会兰, 童康, 孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报, 2019, 47(5): 1162-1173.
LUO Hui-lan, TONG Kang, KONG Fan-sheng. Review of human action recognition in videos based on deep learning [J]. Acta Electronica Sinica, 2019, 47(5): 1162-1173. (in Chinese)

[2] YAN S, XIONG Y, LIN D, et al. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 7444-7452.

[3] SI C, CHEN W, WANG W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE,2019: 1227-1236.

[4] SHI L, ZHANG Y, CHENG J, et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing, 2020, 29: 9532-9545.

[5] LEI T, ZHANG Y, WANG S I, et al. Simple recurrent units for highly parallelizable recurrence[EB/OL]. (2018) [2021]. https://arxiv.org/abs/1709.02755v5.

[6] SHE Q, MU G, GAN H, et al. Spatio-temporal SRU with global context-aware attention for 3D human action recognition[J]. Multimedia Tools and Applications, 2020, 79(17-18): 12349-12371.

[7] PARK C, LEE C, HONG L, et al. S2-Net: Machine reading comprehension with SRU based self-matching networks[J]. ETRI Journal, 2019, 41(3): 371-382.

[8] ZHU W, LAN C, XING J, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix: AAAI,

2016: 3697-3703.

[9] ZHANG L, ZHU G, MEI L, et al. Attention in convolutional LSTM for gesture recognition[C]//Proceedings of the Advances in Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018: 1953-1962.

[10] Di Gangi M A, Federico M. Deep neural machine translation with weakly-recurrent units[EB/OL]. (2018) [2021]. https://arxiv.org/abs/1805.04185.

[11] SONG S, LAN C, XING J, et al. An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017:4263-4270.

[12] 朱红蕾, 朱昶胜, 徐志刚. 人体行为识别数据集研究进展[J]. 自动化学报, 2018, 44(6): 978-1004.
ZHU Hong-lei, ZHU Chang-sheng, XU Zhi-gang. Research advances on human activity recognition datasets[J]. Acta Automatica Sinica, 2018, 44(6): 978-1004. (in Chinese)

[13] XIE C, LI C, ZHANG B, et al. Memory attention networks for skeleton-based action recognition[EB/OL]. (2018)[2021]. https://arxiv.org/abs/1804.08254.

[14] WANG J, NIE X, XIA Y, et al. Cross-view action modeling, learning and recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 2649-2656.

[15] CHENG K, ZHANG Y, HE X, et al. Skeleton-based action recognition with shift graph convolutional Network [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 183-192.

[16] 穆高原. 基于深度学习的危险驾驶行为识别研究[D]. 杭州: 杭州电子科技大学, 2020.
MU Gao-yuan. Study on dangerous driving behavior recognition based on deep learning[D]. Hangzhou: Hangzhou Dianzi University, 2020. (in Chinese)

[17] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 588-595.

[18] WANG J, LIU Z, WU Y, et al. Learning actionlet ensemble for 3D human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(5): 914-927.

[19] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1110-1118.

[20] LIU M, LIU H, CHEN C. Enhanced skeleton visualization for view invariant human action recognition[J]. Pattern Recognition, 2017, 68: 346-362.

[21] LEE D Kim, KANG S, LEE S. Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 1012-1020.

[22] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12026-12035.

[23] SHAHROUDY A, LIU J, NG T T, et al. Nturgb+d: A large scale dataset for 3D human activity analysis[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1010-1019.

作者简介

赵俊男　男, 1996年12月出生于浙江湖州. 现为杭州电子科技大学自动化学院硕士研究生, 研究方向为3D骨架动作识别、人体姿态估计.
E-mail: 663261972@qq.com

佘青山(通讯作者)　男, 1980年2月出生于湖北松滋. 现为杭州电子科技大学教授, 主要研究方向为机器学习与脑-机接口、康复机器人、图像/视频处理与分析.
E-mail: qsshe@hdu.edu.cn