改进通道注意力机制下的人体行为识别网络

陈 莹* 龚苏明

(江南大学轻工过程先进控制教育部重点实验室 无锡 214122)

摘 要：针对现有通道注意力机制对各通道信息直接全局平均池化而忽略其局部空间信息的问题，该文结合人体行为识别研究提出了两种改进通道注意力模块，即矩阵操作的时空(ST)交互模块和深度可分离卷积(DS)模块。ST模块通过卷积和维度转换操作提取各通道时空加权信息数列，经卷积得到各通道的注意权重；DS模块首先利用深度可分离卷积获取各通道局部空间信息，然后压缩通道尺寸使其具有全局的感受野，接着通过卷积操作得到各通道注意权重，进而完成通道注意力机制下的特征重标定。将改进后的注意力模块插入基础网络并在常见的人体行为识别数据集UCF101和HDBM51上进行实验分析，实现了准确率的提升。

关键词：行为识别；通道注意力；时空特征；深度可分离卷积

中图分类号：TN911.73; TP391.4 **文献标识码**：A **文章编号**：1009-5896(2021)12-3538-08

DOI: 10.11999/JEIT200431

Human Action Recognition Network Based on Improved Channel Attention Mechanism

CHEN Ying GONG Suming

(*Key Laboratory of Advanced Process Control for Light Industry* (*Ministry of Education*), *Jiangnan University*, *Wuxi* 214122, *China*)

Abstract: To tackle the problem that the existing channel attention mechanism uses global average pooling to generate channel-wise statistics while ignoring its local spatial information, two improved channel attention modules are proposed for human action recognition, namely the Spatial-Temporal (ST) interaction block of matrix operation and the Depth-wise-Separable (DS) block. The ST block extracts the spatiotemporal weighted information sequence of each channel through convolution and dimension conversion operations, and obtains the attention weight of each channel through convolution. The DS block uses firstly depth-wise separable convolution to obtain local spatial information of each channel, then compresses the channel size to make it have a global receptive field. The attention weight of each channel is obtained via convolution operation, which completes feature re-calibration with the channel attention mechanism. The proposed attention block is inserted into the basic network and experimented over the popular UCF101 and HDBM51 datasets, and the results show that the accuracy is improved.

Key words: Action recognition; Channel attention; Spatiotemporal feature; Depth-wise-Separable(DS) convolution

1 引言

在计算机视觉领域，对人类行为识别的研究既能发展相关理论基础又能扩大其工程应用范围。对于理论基础，行为识别领域融合了图像处理、计算机视觉、人工智能、人体运动学和生物科学等多个学科的知识，对人类行为识别的研究可以促进这些学科的共同进步。对于工程应用，视频中的人类行为识别系统有着丰富的应用领域和巨大的市场价值，其应用领域包括自动驾驶、人机交互、智能安防监控等。

早期的行为识别方法主要依赖较优异的人工设计特征，如密集轨迹特征[1]、视觉增强单词包法[2]等。得益于神经网络的发展，目前基于深度学习的行为识别方法已经领先于传统的手工设计特征的方法。尽管如此，基于深度学习的人体行为识别方法依旧存在着难点：Karpathy等人[3]率先将神经网络

收稿日期：2020-05-29；改回日期：2021-06-03；网络出版：2021-08-24
*通信作者： 陈莹 chenying@jiangnan.edu.cn
基金项目：国家自然科学基金(61573168)
Foundation Item: The National Natural Science Foundation of China (61573168)

运用于行为识别，其将单张RGB图作为网络的输入，这只考虑了视频的空间表观特征，而忽略了时域上的运动信息。Simonyan等人[4]提出了双流网络。该方法使用基于RGB图片的空间流卷积神经网络和基于光流图的时间神经网络分别提取人类行为的静态特征和动态特征，最后将双流信息融合进行识别。一个视频通常持续几秒至几十秒，Wang等人[5]提出了TSN结构来处理此问题，其将一个输入视频分成K段，然后每个段中随机采样得到一个片段。不同片段的类别得分采用段共识函数进行融合来产生段共识。最后对所有模型的预测融合产生最终的预测结果。借鉴2D卷积神经网络在静态图像的成功，Ji等人[6]将2D卷积拓展为3D卷积，从而提出了3D-CNN方法来提取视频中的运动信息。但3D-CNN计算参数太过庞大，难以优化。Zhu等人[7]提出了伪双流结构，网络采用RGB序列作为输入，分支1提取表观信息；分支2则通过图像重建的方法来获得运动信息，然后将预测结果映射到真实标签上。

上述方法都注重寻找额外的时间维信息，如光流运动信息、帧间信息等，而忽略了RGB图像本身富含着重要且丰富的信息。人类在观察不同行为时，对整个空间区域会有不同的关注度，会更加注意人体进行活动的区域。引入空间注意力机制有助于关键特征的增强，提升网络判别性能。Sharma等人[8]首次将注意力机制引入到行为识别中来提升网络在空域上提取关键信息的能力。相比之前的方法，该方法成功提高了识别正确率但结果依旧较低且只关注高层特征。胡正平等人[9]将2维通道注意力拓展为3维通道注意力并运用到3维网络中提升网络的特征提取能力。本文在分析现有通道注意力模块不足的基础上，提出了改进的通道注意力模块，并将此模块插入现有基础网络(如ResNet[10])，实现了识别正确率的提升。

2　注意力机制下的行为识别网络

2.1　现有通道注意力模块

卷积神经网络中，每一张图片初始会由RGB三通道表示出来，之后经过不同的卷积操作，每一个通道又会生成新的信息。每个通道的特征表示的是该输入在不同卷积核上的分量，这些分量对关键信息的贡献有多有少，因此受人类注意力感知机制启发，在网络中加入通道注意力映射模块，能有效建模通道间关系从而提升网络特征提取能力。Hu等人[11]提出了轻量级可插入注意力(Squeeze-and-Excitation, SE)模块，其结构如图1所示。此模块主要构成部分为维度压缩模块、激励、加权。该模块首先利用全局平均池化(global average pooling)操作将每个2维的特征通道变成一个实数，然后利用全连接操作与激活函数(ReLU, Sigmoid)得到比较全面的通道级别的权重关系，最后利用元素乘法将得到的权重与原始特征进行融合。

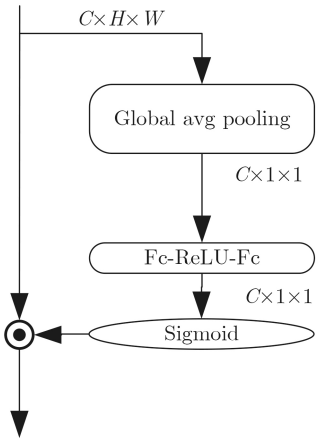


图 1 SE模块

2.2　改进通道注意力模块

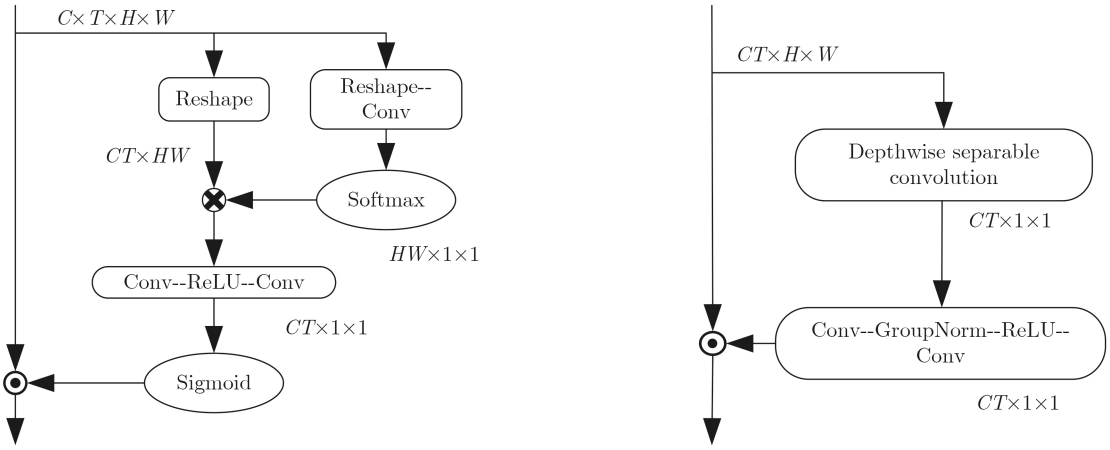
行为识别的主体是人，对于人这个目标来说，中心位置和边界位置的权重应该是不同的。SE_Block中采用全局平均池化操作赋予特征图每个位置相同的权重，在某种程度上加强了不重要信息，抑制了重要信息。为了赋予特征图每个位置可学习权重，本文考虑了两种改进的注意力模块：(1) 矩阵操作的时空交互(Spatial-Temporal, ST)模块，如图2(a)所示；(2) 深度可分离卷积的特征提取(Depth-wise-Separable, DS)模块，如图2(b)所示。

和SE模块一样，本文提出的改进注意力模块是一种即插即用模块，因此可以直接在现有基础网络中加入改进后的注意力模块构成新的识别网络。以DS模块和ResNet为例，图3给出了网络模块示意图。图3(a)为原始ResNet残差块，图3(b)则是加入DS模块之后的网络模块。

3　改进通道注意力模块详解

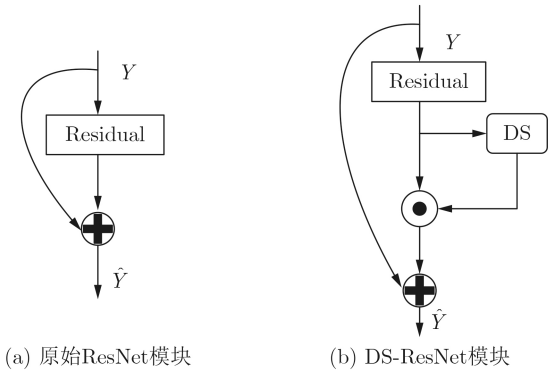
3.1　矩阵操作的时空交互模块

神经网络中的特征图本质上是矩阵数据，那么便可以采用矩阵乘法来对特征图进行处理。在矩阵乘法中需要注意操作数维度匹配问题，综合考虑这些因素，图4给出了ST模块的详细细节。从图中可以知道该模块主要分为3个部分：输入维度转换(图4(b))、时空交互模块(图4(c))、激励加权(图4(d))。为了简化输入，本文将批尺寸(batchsize)省略了。在图4中，输入维度为$[C\times T\times H\times W]$，其中$C$表示通道数，$T$表示图像序列数值，$H$表示高度，$W$表示宽度。若网络输入为单张RGB图像，则$T$为1，若输入为RGB序列，则$T$为序列数值。



(a) ST模块 (b) DS模块

图 2 改进的通道注意力模块



(a) 原始ResNet模块 (b) DS-ResNet模块

图 3 网络模块示意图

在输入维度转换部分，针对模块输入[C×T×H×W]通过简单的矩阵转换操作将其变为[CT×HW]，这样便得到了矩阵乘法的第1个操作数。

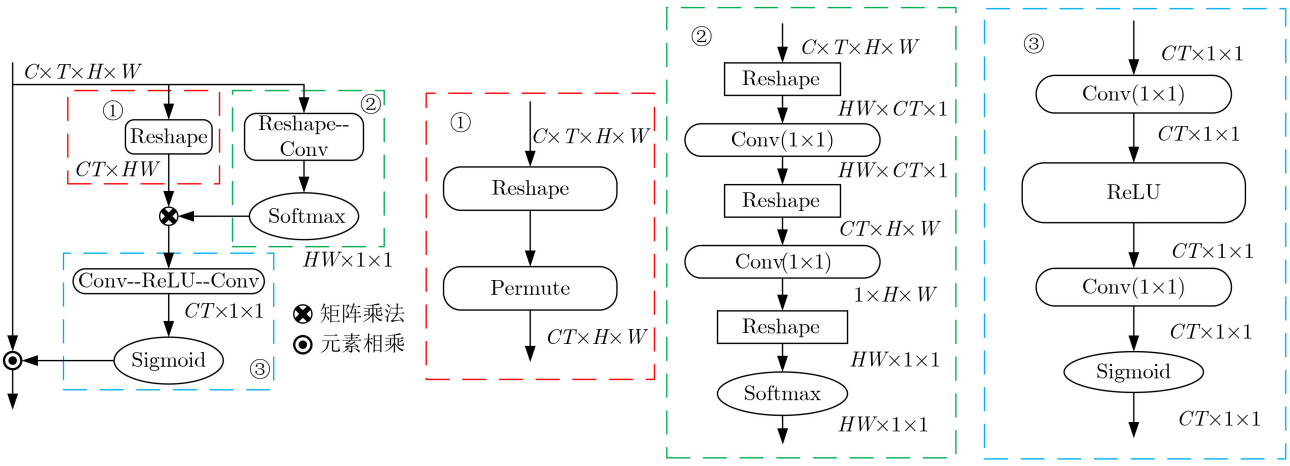
时空交互模块输入与维度转换模块相同，此模块目的是获得矩阵乘法的第2个操作数，同时进行赋权重操作。考虑到输入若为RGB序列，那么在维度变换过程中同时提取输入特征间的相关信息对整体结果会有所提升，因此采用Reshape-Conv复合操作来达到这一目的。2.1节已经提到，特征图每个位置的权重是不同的，于是在此部分结束处使用Softmax操作对每个位置赋予不同的可学习权重。

将上述两部分的输出进行矩阵乘法便得到了第3部分的输入。第3部分采用文献[11]提出的激励加权操作，其作用是将通道权重通过乘法逐通道加权到原来的特征上，完成在通道维度上的原始特征重标定。

3.2 深度可分离卷积的特征提取模块

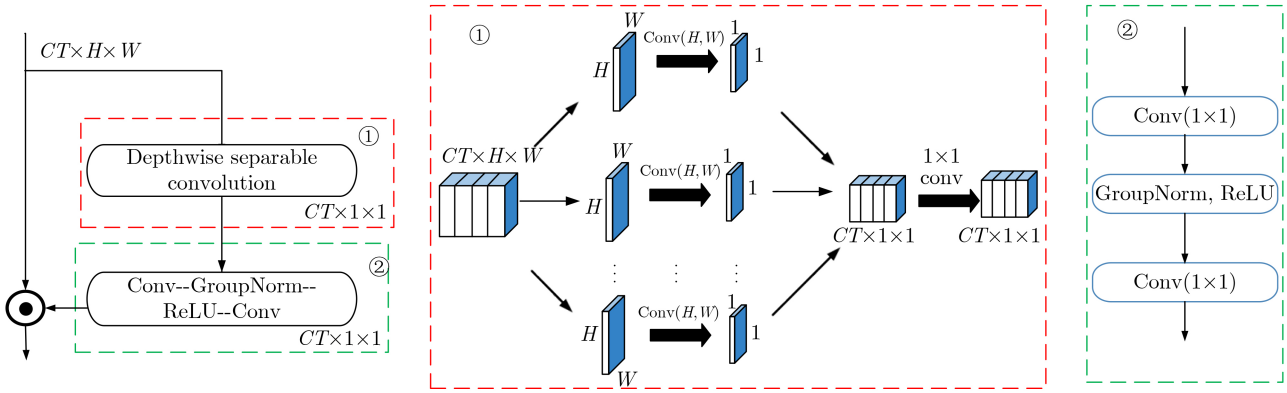
虽然ST模块能满足建模通道间关系这一要求，但其操作复杂且引入的额外计算参数过多，因此提出了DS(图5(a))这一更有效的模块。DS模块主要分为两部分：维度压缩、激励加权。

在维度压缩部分，利用深度可分离卷积来实现，其详细操作见图5(b)。现作出如下假设：输入($C_{\rm in}$×H×W)，卷积核(K_1×K_2)，卷积核数量



(a) ST模块 (b) 维度转换 (c) 时空交互 (d) 激励加权

图 4 ST模块详细示意图



(a) DS模块　　(b) 维度压缩　　(c) 激励加权

图 5 DS_Block详细示意图

(C_{out})，分组数(G)。对于正常卷积，参数数量为：$C_{\text{in}}\times K_1\times K_2\times C_{\text{out}}$；采取分组卷积，参数数量则为：$(1/G)\times C_{\text{in}}\times K_1\times K_2\times C_{\text{out}}$，参数数量比之前减少了$G$倍。当$C_{\text{in}}=C_{\text{out}}=G$时，分组卷积就是Depthwise-Conv。更进一步，当$C_{\text{in}}=C_{\text{out}}=G$，且$K_1=H$，$K_2=W$时，输出特征图尺寸就成了$C_{\text{out}}\times 1\times 1$，实现了全局池化的功能同时赋予了特征图每个位置可学习的权重。

在激励加权部分(图5(c))，相比于SE模块与ST模块，做出了2个改动。首先，由于BatchNorm[12]操作每次计算均值和方差是在一个批量(batch)上，所以如果批尺寸(batchsize)太小，则计算的均值、方差不足以代表整个数据分布，因此采用GroupNorm[13]来替代，这样便与批尺寸(batchsize)无关，不受其约束。当有较好预训练时，可以考虑不使用。其次，考虑到sigmoid函数存在两端饱和，在传播过程中容易丢弃信息，因此可以考虑将其舍弃。

4　损失函数

损失函数选择的是交叉熵函数，其表达式为

$$\text{CE}(p,y)=\begin{cases}-\lg(p), & y=1\\ -\lg(1-p), & \text{其他}\end{cases}\tag{1}$$

其中，p和y分别为预测值与真实标签。

定义一个p_t:

$$\text{p}_\text{t}=\begin{cases}p, & y=1\\ 1-p, & \text{其他}\end{cases}\tag{2}$$

那么，便可以得到

$$\text{CE}(p,y)=\text{CE}(\text{p}_\text{t})=-\lg(\text{p}_\text{t})\tag{3}$$

考虑到可能存在样本不均衡的情况，本文采用了Lin等人[14]提出的Focal Loss函数作为网络的损失函数。Focal Loss函数是交叉熵函数的改进版，其表达式为

$$\text{FL}(\text{p}_\text{t})=-\alpha(1-\text{p}_\text{t})^{\gamma}\lg(\text{p}_\text{t})\tag{4}$$

普通的交叉熵对于正样本而言，输出概率越大损失越小。对于负样本而言，输出概率越小则损失越小。此时的损失函数在大量简单样本的迭代过程中比较缓慢且可能无法优化至最优。因此Focal Loss引入了平衡因子α，其主要用来平衡正负样本。为了解决简单与困难样本的问题，Focal Loss还引入了另一个平衡因子γ。周波等人[15]在其论文中经过实验分析得出，当γ取值范围在2～5之间时，结果一样。因此在本文中，γ取值为2，α取值为0.75。

5　实验与分析

5.1　实验数据集

本文在最常见的行为识别数据集UCF101和HMDB51上对本文网络结构进行评估实验，以便将其性能与目前主流的方法进行比较。

UCF101数据集是从YouTube收集的具有101个动作类别的逼真动作视频的动作识别数据集。101个动作类别中的视频分为 25 组，每组可包含4～7个动作视频。来自同一组的视频可能共享一些共同的功能，例如类似的背景、类似的观点等。

HMDB51数据集内容主要来自电影，一小部分来自公共数据库，如YouTube视频。该数据集包含6849个剪辑，分为51个动作类别，每个动作类别至少包含101个剪辑。

5.2　实验设置

本文实验中，卷积神经网络基于PyTorch平台设计实现。网络训练采用小批量随机梯度下降法，动量为 0.9，权值在每35个epoch衰减一次，衰减率为0.1，损失函数采用Focal Loss函数，其平衡因子α, γ分别为0.75和2。HMDB51数据集的批大小为2，UCF101数据集的批大小为4。本文网络是在ImageNet数据库上预训练的Resnet网络修改而来，初始学习率设为0.01。

5.3 实验结果与分析

5.3.1 注意力模块验证

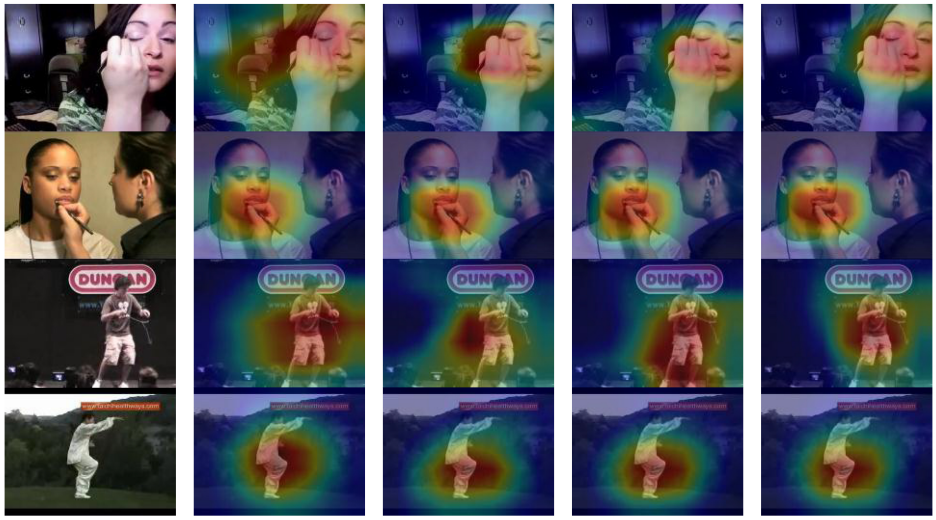
本文重点是注意力机制，因此本节对提出的注意力模块进行验证。首先，图6给出了在ResNet50网络中分别加入SE模块、DS模块、ST模块之后的可视化结果。在图6中，图6(a)表示原图，图6(b)为ResNet50输出结果，图6(c)表示加SE模块后的结果，图6(d)是加ST模块后的结果，图6(e)为加DS模块后的结果。

从结果能看到加了注意力模块后，网络能更关注有效区域。对于第3行结果，相较于DS模块和ST模块，SE模块出现了明显差距，SE模块关注的无效背景区域更大而且并没有重点关注动作幅度较大的手臂区域；在另外3幅图中，SE模块关注的无效区域也更多一点，这表明本文提出的注意力模块更具优势。

表1给出了3个注意力模块在几个主流方法上的实验结果。在该系列实验中，只采用RGB图作为输入，预训练数据集均为ImageNet，主干网络均为ResNet。从表1结果可以看出，3个注意力模块均对网络预测起到了提升作用。例如TSN[5]，基础正确率为85.7%，加了SE模块后正确率提升了0.4%，而DS模块则带来了1.5%的正确率提升。对于MiCT[16]，DS模块提升效果比SE模块高了0.7%。从整体结果来看，DS模块提升最大，ST模块虽然也有提升，但效果不如DS模块。

注意力模块的引入将带来额外的网络参数，对此，本文通过实验给出了各个模块对网络的具体影响，结果如表2所示。对于MiCT[16]网络来说，SE与DS只引入了0.14 M左右的额外参数，这对整个网络而言计算负担并不是很大，而ST模块则引入了较多的额外参数，这对整个网络是不利的。同样，对于ResNet[10]结构，SE与DS引入的额外参数几乎一样，且对网络影响不大，而ST模块依旧引入了大量的额外参数。综合上述分析，DS模块要优于ST模块。

此外，以ResNet50为Baseline，对增加注意力模块前后的精度和运行时间进行了比较，结果如



(a) 原图 (b) 原始ResNet50 (c) ResNet50+SE (d) ResNet50+ST (e) ResNet50+DS

图 6 不同注意力模块可视化结果

表 1 验证注意力模块

方法	主干网络	UCF101准确率(%)	HMDB51准确率(%)
TSN[5]	ResNet-101	85.7	54.6
TSN+SE	ResNet-101	86.1	55.6
TSN+DS	ResNet-101	87.2	56.4
TSN+ST	ResNet-101	87.0	55.8
MiCT[16]	ResNet-34	69.0	40.5
MiCT+SE	ResNet-34	70.1	41.2
MiCT+DS	ResNet-34	70.8	41.8
MiCT+ST	ResNet-34	70.4	41.3

表3所示。分析比较发现，相比未加入注意力模块的Baseline，加入SE模块后精度提升了0.4%，运行时间增加了0.27 s，加了DS模块后精度提升了1.5%，运行时间增加了1.27 s，加入ST模块后精度提升了1.3%，运行时间增加了2.46 s，DS模块无论在准确率还是速度上都优于ST模块。相比于SE模块，DS模块精度提升增加了1.1%，运行时间增加了1 s。从上述数据可以看出，增加注意力模块都会在提高精度的同时，降低计算速度，而DS模块相比于SE模块，在准确率上取得了较大的提升，但同时增加了一定的计算损耗，今后工作将围绕如何更好地平衡速度精度问题进一步展开。

5.3.2 与主流网络对比结果

通过5.3.1节的验证实验可以知道本文提出的通道注意力模块在提升模型识别精度上的有效性，其中DS模块效果最好。为了与主流网络进行比较，本文将DS模块及Focal Loss运用到TSN, MiCT两个基本网络中，实验结果见表4。

首先，为了验证DS模块对于RGB图像能起到提升作用，在表4上半部分系列实验中，所有网络输入均采用RGB序列。在这些方法中，P3D[17]通过将3D卷积分解成沿空间维度的2D卷积和沿时间维度的1D卷积来构建时空信息模型；I3D[18]直接将最新的2D卷积架构膨胀成3D卷积，以利用预训练的2D模型；TS+LSTM[19]利用2D网络提取视频帧的表征信息，紧接着连接一个循环神经网络(LSTM)来学习帧与帧之间的运动信息。从表4可以看出，加了本文注意力模块的方法相比于其它方法，获得了更好的性能。以UCF101结果为例，DS模块与Focal Loss给TSN[5]带去了1.6%的增长，同时在所有方法中表现最优。此外，TLE[20]采用精心设计的网络结构(BN-Inception)；P3D[17]则使用了更大的预训练数据集(Kinetics)，它们均比原始的TSN[5], MiCT[16]表现更好，但加入DS模块后，后者表现更好，这意味着本文注意力模块对识别结果有较大的提升。

为了验证本文注意力模块对于光流也有效同时方便与其他方法比较，对所有网络采用RGB和光流两种模态输入，实验结果见表5。以UCF101结果为例，对于MiCT[16]方法，构建了RGB流和光流两条支流，表5中MiCT-A表示在RGB流引入DS模块而光流支流中不加入DS模块，其最终识别结果为94.2%；MiCT-B表示在RGB流与光流两流中均引

表 2　网络参数对比结果

方法	主干网络	参数大小(M)
MiCT	ResNet-34	26.16
MiCT+SE	ResNet-34	26.30
MiCT+DS	ResNet-34	26.31
MiCT+ST	ResNet-34	30.09
ResNet-50	ResNet-50	23.71
ResNet-50+SE	ResNet-50	26.20
ResNet-50+DS	ResNet-50	26.22
ResNet-50+ST	ResNet-50	86.61

表 3　注意力模块的精度及运行时间比较

方法	准确率(%)	平均运行时间(s)
ResNet-50	85.7	0.93
ResNet-50+SE	86.1	1.20
ResNet-50+DS	87.2	2.20
ResNet-50+ST	87.0	3.39

表 4　不同算法在UCF101与HMDB51数据集上识别准确率对比(单流输入)

方法	输入	主干网络	预训练	UCF101(%)	HMDB51(%)	fps
C3D[21]	RGB	3D Conv.	Sports-1M	44.0	43.9	4.2
TS+LSTM[19]	RGB	ResNet+LSTM	ImageNet	82.0	–	–
TSN[5]	RGB	ResNet101	ImageNet	85.7	54.6	8.5
LTC[22]	RGB	ResNet-50	ImageNet	83.0	52.8	–
TLE[20]	RGB	3D Conv.	ImageNet	86.3	63.2	–
TLE[20]	RGB	BN-Inception	ImageNet	86.9	63.5	–
I3D[18]	RGB	BN-Inception	ImageNet+Kinetics	84.5	49.8	8.3
P3D[17]	RGB	ResNet-101	ImageNet+Kinetics	86.8	–	13.4
MiCT[16]	RGB	ResNet-101	ImageNet+Kinetics	86.1	62.8	4.8
C-LSTM[8]	RGB	ResNet+LSTM	ImageNet	84.96	41.3	8.0
TSN+DS	RGB	ResNet-101	ImageNet	**87.3**	**64.4**	**3.6**
MiCT+DS	RGB	ResNet-101	ImageNet	**87.0**	**64.2**	**2.1**

表 5 不同算法在UCF101与HMDB51数据集上识别准确率对比(双流输入)

方法	输入	主干网络	预训练	UCF101(%)	HMDB51(%)
DTPP[23]	RGB+FLOW	ResNet-101	ImageNet	89.7	61.1
TS+LSTM[19]	RGB+FLOW	ResNet+LSTM	ImageNet	88.1	–
LTC[22]	RGB+FLOW	ResNet-50	ImageNet	91.7	64.8
TLE[20]	RGB+FLOW	BN-Inception	ImageNet+Kinetics	95.6	70.8
T3D[24]	RGB+FLOW	ResNet-50	ImageNet+Kinetics	91.7	61.1
I3D[18]	RGB+FLOW	BN-Inception	ImageNet	93.2	69.3
TSM[25]	RGB+FLOW	ResNet-50	ImageNet+Kinetics	94.5	70.7
MiCT-A	RGB+FLOW	ResNet-101	ImageNet	**94.2**	**70.0**
MiCT-B	RGB+FLOW	ResNet-101	ImageNet	**94.6**	**70.9**

入DS模块，其最终结果达到了94.6%。这表明本文的注意力模块在光流中依旧能起作用。

6 结论

本文提出了改进注意力机制下的人体行为识别方法。通过分析现有通道注意力机制的不足，提出了改进的注意力模块。为了验证改进注意力模块的有效性，分别从可视化结果、网络精度提升、额外网络参数等方面进行实验验证。最后将模块运用到现有的基础网络中，在通用数据集上与其他主流方法进行比较，实验结果再次证明了改进后的模块的有效性。今后工作将围绕如何提高模块速度上进一步展开。一些说明及源码见：https://github.com/gongsuming/paper1。

参 考 文 献

[1] IKIZLER-CINBIS N and SCLAROFF S. Object, scene and actions: Combining multiple features for human action recognition[C]. The 11th European Conference on Computer Vision, Heraklion, Greece, 2010: 494–507.

[2] 张良, 鲁梦梦, 姜华. 局部分布信息增强的视觉单词描述与动作识别[J]. 电子与信息学报, 2016, 38(3): 549–556. doi: 10.11999/JEIT150410.
ZHANG Liang, LU Mengmeng, and JIANG Hua. An improved scheme of visual words description and action recognition using local enhanced distribution information[J]. *Journal of Electronics & Information Technology*, 2016, 38(3): 549–556. doi: 10.11999/JEIT150410.

[3] KARPATHY A, TODERICI G, SHETTY S, *et al*. Large-scale video classification with convolutional neural networks[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Colombus, USA, 2014: 1725–1732.

[4] SIMONYAN K and ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. The 27th International Conference on Neural Information Processing Systems - Volume 1, Montreal, Canada, 2014: 568–576.

[5] WANG Limin, XIONG Yuanjun, WANG Zhe, *et al*. Temporal segment networks: Towards good practices for deep action recognition[C]. The 14th European Conference, Amsterdam, The Kingdom of the Netherlands, 2016: 20–36.

[6] JI Shuiwang, XU Wei, YANG Ming, *et al*. 3D convolutional neural networks for human action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1): 221–231. doi: 10.1109/TPAMI.2012.59.

[7] ZHU Yi, LAN Zhenzhong, NEWSAM S, *et al*. Hidden two-stream convolutional networks for action recognition[C]. The 14th Asian Conference on Computer Vision, Perth, Australia, 2018: 363–378.

[8] SHARMA S, KIROS R, and SALAKHUTDINOV R. Action recognition using visual attention[C]. The International Conference on Learning Representations 2016, San Juan, The Commonwealth of Puerto Rico, 2016: 1–11.

[9] 胡正平, 刁鹏成, 张瑞雪, 等. 3D多支路聚合轻量网络视频行为识别算法研究[J]. 电子学报, 2020, 48(7): 1261–1268. doi: 10.3969/j.issn.0372-2112.2020.07.003.
HU Zhengping, DIAO Pengcheng, ZHANG Ruixue, *et al*. Research on 3D multi-branch aggregated lightweight network video action recognition algorithm[J]. *Acta Electronica Sinica*, 2020, 48(7): 1261–1268. doi: 10.3969/j.issn.0372-2112.2020.07.003.

[10] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016: 770–778.

[11] HU Jie, SHEN Li, and SUN Gang. Squeeze-and-excitation networks[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7132–7141.

[12] IOFFE S and SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB/OL]. https://arxiv.org/abs/1502.03167v2, 2015.

[13] WU Yuxin and HE Kaiming. Group normalization[C]. The

European Conference on Computer Vision, Amsterdam, The Kingdom of the Netherlands, 2018: 3–19.

[14] LIN T Y, GOYAL P, GIRSHICK R, *et al*. Focal loss for dense object detection[C]. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 2999–3007.

[15] 周波, 李俊峰. 结合目标检测的人体行为识别[J]. 自动化学报, 2020, 46(9): 1961–1970.
ZHOU Bo and LI Junfeng. Human action recognition combined with object detection[J]. *Acta Automatica Sinica*, 2020, 46(9): 1961–1970.

[16] ZHOU Yizhou, SUN Xiaoyan, ZHA Zhengjun, *et al*. MiCT: Mixed 3D/2D convolutional tube for human action recognition[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 449–458.

[17] QIU Zhaofan, YAO Ting, and MEI Tao. Learning spatio-temporal representation with pseudo-3D residual networks[C]. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 5534–5542.

[18] CARREIRA J and ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017: 4724–4733.

[19] MA C Y, CHEN M H, KIRA Z, *et al*. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition[J]. *Signal Processing*: *Image Communication*, 2019, 71: 76–87. doi: 10.1016/j.image.2018.09.003.

[20] DIBA A, SHARMA V, and VAN GOOL L. Deep temporal linear encoding networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017: 1541–1550.

[21] TRAN D, BOURDEV L, FERGUS R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C]. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015: 4489–4497.

[22] VAROL G, LAPTEV I, and SCHMID C. Long-term temporal convolutions for action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(6): 1510–1517. doi: 10.1109/TPAMI.2017.2712608.

[23] ZHU Jiagang, ZHU Zheng, and ZOU Wei. End-to-end video-level representation learning for action recognition[C]. 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 2018: 645–650.

[24] DIBA A, FAYYAZ M, SHARMA V, *et al*. Temporal 3D convnets: New architecture and transfer learning for video classification[EB/OL]. https://arxiv.org/abs/1711.08200v1, 2017.

[25] LIN Ji, GAN Chuang, and HAN Song. TSM: Temporal shift module for efficient video understanding[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea(south), 2019: 7082–7092.

陈　莹：女，1976年生，教授，博士，研究方向为信息融合、模式识别. Euclid.

龚苏明：男，1995年生，硕士生，研究方向为计算机视觉与模式识别.

责任编辑：马秀强