DOI:10.19651/j.cnki.emt.2209232

基于 2D CNN 和 Transformer 的人体动作识别*

朱相华 智 敏 殷雁君
(内蒙古师范大学计算机科学技术学院 呼和浩特 010022)

摘 要:人体动作识别是计算机视觉领域的研究热点之一,在人机交互、视频监控等方面具有深远的理论研究意义。为了解决 2D CNN 无法有效获取时间关系等问题,利用 Transformer 在建模长期依赖关系上的优势,引入 Transformer 架构并将其与 2D CNN 相结合用于人体动作识别,以更好地捕获上下文时间信息。首先使用融合通道-空间注意力模块的 2D CNN 提取强化的帧内空间特征,其次利用 Transformer 捕捉帧间的时间特征,最后应用 MLP Head 进行动作分类。实验结果表明在 HMDB-51 数据集和 UCF-101 数据集上分别达到了 69.4%和 95.5%的识别准确度。

关键词:人体动作识别;2D CNN;通道-空间注意力模块;Transformer
中图分类号:TP18 **文献标识码**:A **国家标准学科分类代码**:520.1

Human action recognition based on 2D CNN and Transformer

Zhu Xianghua Zhi Min Yin Yanjun
(College of Computer Science and Technology, Inner Mongolia Normal University, Hohhot 010022, China)

Abstract: Human action recognition is one of the research hot-spots in the field of computer vision. It has far-reaching theoretical research significance in human-computer interaction, video surveillance and so on. In order to solve the problem that 2D CNN can not effectively obtain time relationship, based on the advantages of Transformer in modeling long-term dependency, Transformer structure is introduced and combined with 2D CNN for human action recognition to better capture context time information. Firstly, 2D CNN integrating channel-spatial attention module is used to capture the inter spatial features. Then, Transformer is used to capture the temporal feature between frames. Finally, MLP head is used for action classification. The experimental results show that the recognition accuracy of HMDB-51 datasets and UCF-101 datasets is 69.4% and 95.5% respectively.

Keywords: human action recognition; 2D CNN; channel-spatial attention module; Transformer

0 引 言

近年来,随着抖音、快手、西瓜视频等各种视频软件的广泛应用,每天都有大量的视频上传到网络,视频流呈现爆炸性增长的趋势。人们对视频内容调节(即识别视频中违反服务条款的内容)和内容推荐(即视频按最喜欢的内容排序并推荐给类似的用户)有了更高要求。如何对视频进行有效的处理和分析逐渐成为关注的热点,使得视频理解应运而生。人体动作识别作为视频理解的重要研究方向也受到大众的广泛关注。视频中复杂动作的相关性包括每个帧内的空间信息和帧间的时间信息,主要表现在伴随着时间帧的向前移动,视频中动作的复杂程度也在不断变化。例如"打开盒子"与"关闭盒子"在空间域中均是类似的特征,时间信息恰恰完全相反。传统的人体动作识别更多地与场景相关[1-3],但是随着手势与环境进行交互的在线识别技术的快速发展,与时间相关的人体动作识别也成为当下的研究热点。

2D CNN 是目前人体动作识别的主流方法之一,基于 2D CNN 的框架不仅具有轻量级和快速推理能力的优点,还可以对整个视频进行稀疏采样的短片段进行操作。但是 2D CNN 仍存在对部分动作特征表达不足和缺乏时间建模的能力。

针对问题 1,为了捕获视频中包含的多种类型的信息,在提取空间信息时,本文设计了一个简单而有效的基于 2D

收稿日期:2022-03-11
*基金项目:内蒙古自治区高等学校科学研究项目(NJZZ21004)、内蒙古师范大学研究生科研创新基金(CXJJS21159)、内蒙古自然科学基金(2018MS06008)项目资助

CNN 的注意力机制，将通道注意力（channel attention module，CAM）和空间注意力（spatial attention module，SAM）嵌入到 2D CNN 架构中，以最小的额外计算成本提取出更多的特征。CAM 是一种新颖的通道注意力模块，其灵感来自于 SE Block[4]，用于区分序列中的关键帧，通过明确地建模通道之间的相互依赖关系自适应地重新校准通道特征。SAM 是空间注意力模块，根据对识别结果的贡献，将不同程度的注意力分配到特征图的不同位置，使动作聚焦于人体动作的运动区域。

针对问题 2，受 Transformer 架构[5]在自然语言处理（nature language process，NLP）领域成功推动的启发，研究人员正试图将 Transformer 应用于视觉任务，Transformer 是建立在多头自注意力层上，学习对序列中元素的全局关注，具有长期依赖性的特点。在目标检测[6]、图像分割[7]、实例分割[8]等领域，实现了与 2D CNN 相当甚至更好的性能。为此，本文将视觉 Transformer[9]自然扩展到视频，将视频采样帧视为文本段落的模拟，强调在人体动作识别任务中提取动作时间信息的重要性。以此改进 2D CNN 缺乏时间建模的能力，使得“顺时针旋转”与“逆时针旋转”等动作避免丢失重要的序列信息。因此，本文提出了将 2D CNN 和 Transformer 组成动作识别模型，该模型利用改进的通道-空间注意力机制提取帧内的空间特征，利用时间 Transformer 提取帧间的时间特征，并将其命名为 ConvTransformer。主要贡献如下：

1）将 Transformer 架构引入到人体动作识别研究领域，使视频时间信息得到了充分的理解。

2）构造了改进的通道注意力机制，并与空间注意力机制级联，以提高识别精度。

3）提出了一种新的基于视频的人体动作识别方法 ConvTransformer，在 ConvTransformer 中融合 2D CNN 架构和 Transformer 模型，最大限度的发挥了卷积和自注意力机制的优势。在 HMDB-51 数据集和 UCF-101 数据集上分别达到了 69.4%和 95.5%的动作识别准确率。

1　相关研究

ConvTransformer 通过改进的通道-空间注意力机制的 2D CNN 获取空间信息，利用 Transformer 获取时间信息。所以在本节中，考虑从 2D CNN 框架、注意力机制和 Transformer 架构 3 个角度分别讨论相关工作。

1.1　2D CNN 框架

基于 2D CNN 的方法独立处理每个帧，以提取视频帧的特征，然后通过在网络末端执行时间平均建模进行特征聚合。Two-Stream[10]在大型数据集上探索了包含空间网络和时间网络的双流 CNN，以学习运动和外观特征。在此基础之上，TSN[11]将视频进行“分段”处理，在长视频序列上利用稀疏采样的方式获取短片段。Trajectory Pooling[12]是有效的线性池方法，沿着时间线堆叠特征，进行动作识别。STCNN[13]将空间仿射变换网络（spatial transform network，STN）和 CNN 架构相结合，在时间上采用了多帧稠密光流算法，并且采用时间与空间相同的 CNN 架构，然后再使用加权求和的方法对时空特征进行融合；LTC[14]探索了具有长期时间卷积的视频表示，以在完整的时间范围内模拟动作。TSM[15]首先将时间建模技术融入到了基于 2D CNN 的框架中，将一部分通道信息通过移位操作内嵌到 2D CNN 中。但是这些工作直接使用 2D CNN，导致并没有明确的的动作时间建模，如相邻帧之间的差异。最近，有几个工作建议把注意力模块内嵌在二维 CNN 中，对空间信息和时间信息进行建模，例如 STACNet[16]通过融合特征图的价值特征和梯度特征建立空间注意模型，使动作识别的卷积集中在动作的信息运动区域；并利用 TAM 模块探索视频中的关键帧。虽然该方法在模拟短时间距离视频序列中具备较高的性能，但是在捕获长时间上下文信息时效果还有待提高。

1.2　注意力机制

为了探索视频帧中动作识别的相关区域并对其给予更高的权重，STA-LSTM[17]介绍了一种基于动作识别骨架数据的端到端时空注意力机制，包括空间注意力模块将不同的注意力分配给每个帧内输入骨架的不同关节，时间注意力模块旨在区分序列中的关键帧。考虑到用于动作识别的注意力模块中的原始 LSTM 没有明确考虑每个骨骼关节相对于全局动作序列的信息性，GCA-LSTM[18]利用全局上下文信息进行注意，有选择性地将更多的注意力放在每个帧中的信息关节上。通过将滑动窗口内的相邻帧与注意力权重进行融合。Multi-LSTM[19]是一种基于 LSTM 的新型深度网络，用于建模类内和类间的时间关系。最近，Li 等[20]和 Tang 等[21]展示了强化学习算法可以有效地选择视频中的重要帧和关键帧中要运动的相关区域。然而，上述所有注意模块都基于 RNN，使得模型硬件效率低下，难以训练。显然，应该更加努力地开发基于 CNN 的注意力模块和 Transformer 模块，并表现出与 RNN 不同的特点。

1.3　Transformer 架构

视觉 Transformer 的首次工作是直接将 Transformer 架构应用于非重叠的图像块，用于图像分类。同卷积神经网络相比，它在图像分类上实现了速度和精度之间良好的折衷。虽然 ViT 需要大规模的训练数据集 JFT-300M 才能很好地执行，但 DeiT[22]引入了几种训练策略，使 ViT 也能有效地使用较小的 ImageNet-1K 数据集。由于视觉 Transformer 的巨大成功导致了对基于视频识别任务 Transformer 体系结构的研究，VTN[23]建议在预先训练的 ViT 之上添加一个时间注意力编码器，在视频动作识别方面产生了良好的性能。TimeSformer[24]研究了时空注意的 5 种不同变体，并提出了一种因式分解的时空注意力，使得无论在速度还是准确性上都有很强的权衡。ViViT[25]研究了预先训练的 ViT 模型的 4 种空间和时间注意力的因子

化设计,并提出了一种类似于 VTN 的架构,该架构在 Kinetics 数据集上实现了最先进的性能。MViT[26]是一种用于视频识别的多尺度视觉 Transformer,通过集中注意力进行时空建模来减少计算量,从而在 Something-SomethingV2 上获得最先进的结果。以上模型无论是在空间还是时间都完全基于 Transformer 模型,但是产生了大量的参数,增加了计算负担。然而,本文尝试将 CNN 和 Transformer 相结合,利用各自优势进行动作识别,主要分为空间和时间两个维度,在空间上依赖于 CNN 架构获得局部信息,在时间上依赖 Transformer 模型获得全局信息。

2 本文方法

本文的目标是提供一个基于改进的 2D CNN 架构来提取空间信息和基于 Transformer 架构来捕获时间信息的长期依赖关系模型 ConvTransformer,利用稀疏的二次采样时间数据提供准确的动作预测。整体框架图如图 1 所示,首先从视频中均匀采样 16 帧作为 2D CNN 架构的输入,在 2D CNN 架构中引入改进的通道注意力模块 CAM 和空间注意力模块 SAM 提取强化的空间特征,将其得到的特征图经过 1×1 卷积层进行降维;其次利用线性变换层将二维的特征图嵌入转换为长度为 D 的一维向量,并添加位置嵌入信息(0,1,2,…),即每个图像的位置信息,更有利于模型准确的评测注意力权重;再将其输入到时间 Transformer 模块,提供视频时间上下文信息的表示;最后经过 MLP Head 输出人体动作类别预测。

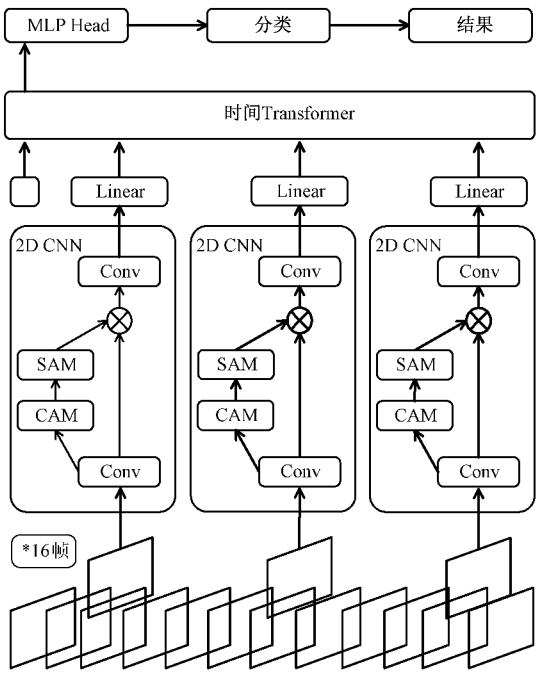

图 1 本文提出的 ConvTransformer 结构

通过模块化设计的 ConvTransformer,将提取的空间特征和时间特征分离有如下优点:首先,在第 1 阶段利用融合改进的通道-空间注意力机制的 CNN 提取强化的帧间空间特征,大大减少了计算量;其次,在第 2 阶段利用 Transformer 提取时间特征,既捕获了全局感受野,又获得了长时间的动作互动信息,使视频信息得到了更好的表达。

2.1 空间特征提取

视频帧中的部分空间特征与时间信息关系不大,如图 2 所示(前 4 张图为"购物",后 4 张图为"采访"),这些动作很容易通过外观特征识别。因此,为了提取更加有效的空间特征,本节在传统的 2D CNN 架构上加入通道-空间注意力模块,主要分为改进的通道注意力模块 CAM 和空间注意力模块 SAM。

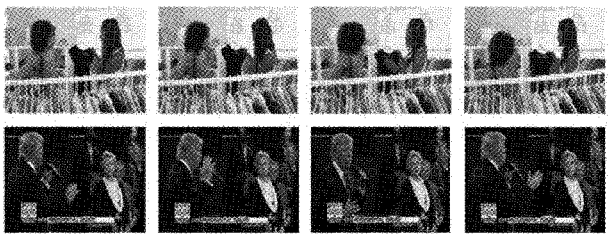

图 2 "购物"与"采访"的轨迹特征

1)改进的通道注意力

Hu 等[4]介绍了通道注意力的新型结构单元(squeeze-and-excitation,SE),自适应地探索神经网络中不同通道的重要性,来提取强化的动作特征。SE 模块的网络结构如图 3 所示,主要分为压缩和激励两个模块:压缩模块是在特征图上执行全局池化操作,将全局信息压缩为特征向量;激励模块是经过两层全连接操作,进而得到特征图中每个通道的权重。特征向量根据通道权重值被重新扩展为特征图,并作为下一层网络的输入。SE 模块的表达式如下:

$$F_{out} = F_{in} \cdot \sigma(P_g(F_{in})) \tag{1}$$

其中,$P_g()$ 表示全局池化,σ 表示 *Sigmoid* 激活函数。

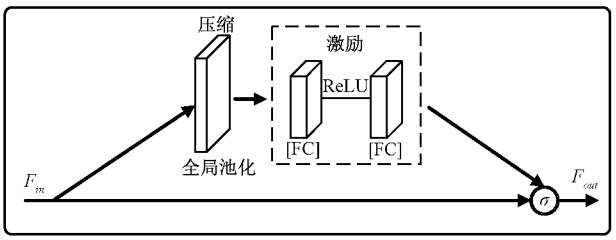

图 3 通道注意力模块(SE)

本文在 SE 模块的基础上,将全局池化改为最大池化和平均池化:平均池化可以在一定程度上保持全局信息的不变性,最大池化可以增强对关键通道的关注度,二者以并行的方式进行操作,构造了改进的通道注意力模块 CAM,网络结构如图 4 所示。在 CAM 模块中,平均池化在向前、向后的传递过程中没有考虑局部信息的重要性,使得贡献度逐渐衰减;由于最大池化相比于平均池化增强了关键通道的关注度,所以在 CAM 模块中本文同时使用平均池化和最大池化,得到更加全面的强化特征图。

由式(2),CAM 模块可以改写为:

$$M_c(F') = \sigma(Avgpool(F) + MaxPool(F)) \tag{2}$$

其中,$AvgPool()$ 表示平均池化,$MaxPool()$ 表示最大池化,σ 表示 $Sigmoid$ 激活函数。

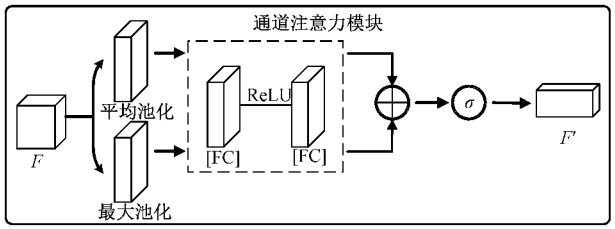

图 4　改进的通道注意力模块(CAM)

2)空间注意力

ConvTransformer 模型融合了改进的通道注意力模块后,强化了关键通道特征,提升了动作识别精度,但对关键结构特征的利用仍然不足。基于这一观察结果,本文在空间维度上同时加入空间注意力机制(SAM),结构图如图 5 所示。该模块以每个特征映射的信息部分"哪里"为核心,根据不同的空间区域对识别结果的贡献,给不同的空间区域赋予不同的权重。

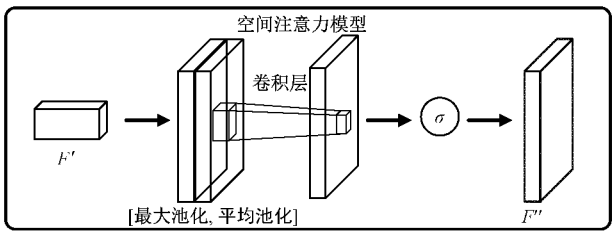

图 5　空间注意力模块(SAM)

为了计算空间注意力,本文首先对通道轴进行平均池化和最大池化操作,再将它们以级联的方式生成有效的特征图。在此基础上,应用卷积层生成空间注意力图 $M_s(F) \in R^{H\times W}$,对突出或模糊的位置进行编码。空间注意力的计算如下:

$$M_s(F'') = \sigma(f^{7\times7}([AvgPool(F');MaxPool(F')])) \tag{3}$$

其中,σ 表示 $Sigmoid$ 激活函数,$f^{7\times7}$ 表示大小为 7×7 的卷积核,$AvgPool()$ 表示平均池化操作,$MaxPool()$ 表示最大池化操作,(;)表示级联操作。本文在设计该模块时,将 CAM 模块与 SAM 模块级联在一起,组成改进的通道-空间注意力模块。

2.2　时间特征提取

基于视频的人体动作识别中动作变换大多与时间相关,不同动作间的判定方法也有所区别。如图 6 所示(前 4 张图为"打开",后 4 张图为"着陆"),"打开"与"着陆"更依赖于时间信息,二者在空间域上的特征差异不大,但在时间域上两者表达的信息完全不同。因此,为了改进 2D CNN 不能很好地提取视频时间序列信息这一缺点,本文将 Transformer 架构引入基于视频的人体动作识别研究中,利用 Transformer 在建模时间信息的长期依赖关系时比卷积更有优势的特点,来更好地处理视频时间信息。

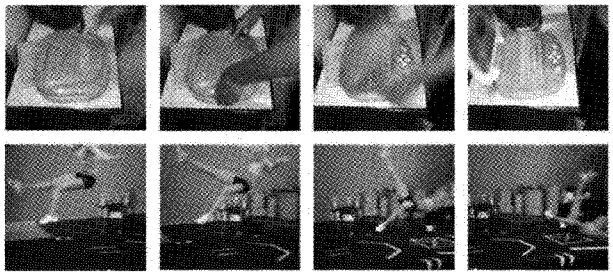

图 6　"打开"与"着陆"的轨迹特征

1)嵌入模块

时间 Transformer 是用于提取视频时间层面特征的模块,以改进的 2D CNN 输出特征图作为输入,将特征图的二维空间特征 F'' 经过线性变换层表示成长度为 D 的一维向量并聚合在一起,作为时间 Transformer 的嵌入,公式如下:

$$\boldsymbol{z}_{(t)}^{(0)} = \boldsymbol{E}\boldsymbol{x}_{(t)} + \boldsymbol{e}_{(t)}^{pos} \tag{4}$$

其中,输入向量 $\boldsymbol{x}_{(t)} \in R$,嵌入向量 $\boldsymbol{z}_{(t)} \in R^D$ 和一个可学习的位置嵌入向量 $\boldsymbol{e}_{(t)}^{pos}$,矩阵 $\boldsymbol{E}$,参数 $t = 1,\cdots,F$ 表示帧的数量。为了使用 Transformer 模型进行分类,在嵌入序列的第 1 个位置添加可学习的分类标记 $z_{(0)}^{(0)} \in R^D$,该分类块将用于编码来自每个帧的信息,并在时间上将其传播到帧序列中。

2)多头自注意力块

时间 Transformer 包含 L 个多头自注意力块(multi-head self-attention block,MSA),结构如图 7 所示,其中每个多头自注意力块 $l \in \{1,\cdots,L\}$、多头 $a \in \{1,\cdots,\Lambda\}$,每个帧转换成 Query、Key、Value 向量,其公式如式(5)~(7)所示,LN 表示线性变换层,每个注意力头的维度是 $D^h = D/\Lambda$。

$$q_{(t)}^{(l,a)} = W_Q^{(l,a)} LN(z_{(t)}^{(l-1)}) \in R^{D_h} \tag{5}$$

$$k_{(t)}^{(l,a)} = W_K^{(l,a)} LN(z_{(t)}^{(l-1)}) \in R^{D_h} \tag{6}$$

$$v_{(t)}^{(l,a)} = W_V^{(l,a)} LN(z_{(t)}^{(l-1)}) \in R^{D_h} \tag{7}$$

首先,Query、Key、Value 向量通过线性变换层再输入到放缩点积注意力中,并且在此部分中需要做 h 次操作。其中,每次 Query、Key、Value 向量进行线性变换的参数 W 不同;然后将放缩点击注意力进行拼接,再进入线性变换层得到的值作为多头自注意力的结果。最后,注意力权重通过 Query 和 Key 之间点积计算,每一帧的注意力权重由式(8)计算:

$$\alpha_{(t)}^{(l,a)} = SM\left(\frac{q_{(t)}^{(l,a)^{\mathrm{T}}}}{\sqrt{D_h}} \cdot \left[k_{(0,t)}^{(l,a)}\{k_{t'}^{(l,a)}\}_{t'=1,\cdots,F}\right]\right) \tag{8}$$

其中,$SM()$ 表示分类激活函数,$k_{(0,t)}^{(l,a)}$ 表示的是第 t 帧的分类块 Key 值。

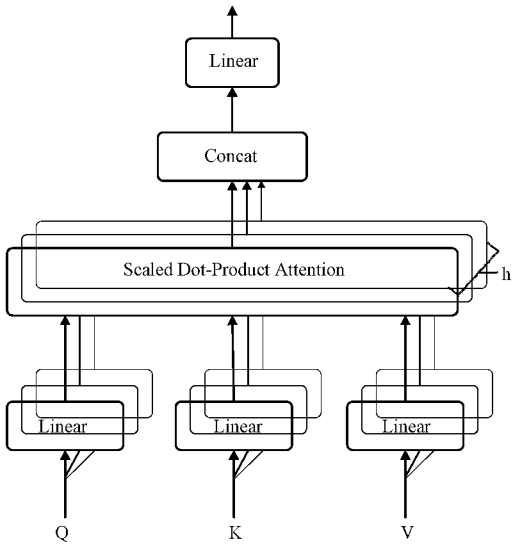

图 7 多头自注意力块

3)时间 Transformer

在对视频时间上下文信息关系进行建模方面,Transformer 模型具备比卷积等同类架构更多的优势。具有足够多头的自注意力层不仅与卷积层一样具有表现力[27],还具备直接模拟远距离交互的能力。因此,为了改进 2D CNN 无法有效获取时间信息的问题,则利用时间 Transformer 进行二次采样,给予视频中的时间上下文信息更好的理解,模型如图 8 所示。

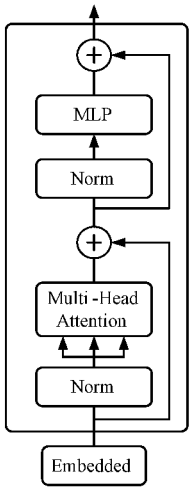

图 8 时间 Transformer 模型

对于模型的时间块部分,本文只在帧索引上计算时间注意力,具体计算过程如下:

$$\alpha_t^{(l,a)time}=SM\left(\frac{q_t^{(l,a)^{\mathrm{T}}}}{\sqrt{D_h}}\cdot\left[k_0^{(l,a)}\left\{k_{t'}^{(l,a)}\right\}_{t'-1,\cdots,F}\right]\right) \tag{9}$$

然后,对每个自注意力头进行加权求和:

$$s_{(t)}^{(l,a)}=\alpha_{(0)}^{(l,a)}v_{(0)}^{(l,a)}+\sum_{t'=1}^{F}\alpha_{(t')}^{(l,a)}v_{(t')}^{(l,a)} \tag{10}$$

再将来自多注意头的输出连接起来,并通过一个带有 GeLU 激活函数的 2 层多层感知器(MLP):

$$z'^{(l)}_{(t)}=W_O\begin{bmatrix}s_{(t)}^{(l,1)}\\ \vdots\\ s_{(t)}^{(1,A)}\end{bmatrix}+z_{(t)}^{(l-1)} \tag{11}$$

$$z_{(t)}^{(l)}=MLP(LN(z'^{(l)}_{(t)}))+z'^{(l)}_{(t)} \tag{12}$$

由于添加了残差连接,所以将 MSA 和 MLP 层进行残差操作。

最后输出时间特征序列:

$$y=LN(z_{(0)}^{(L_{time})})\in R^D \tag{13}$$

其中,L_{time} 是时间注意力层的数量,MLP 被用作分类器,输出一个等于类数的维度向量。

2.3 分 类

将 Transform 输出的时间特征序列的类别块输入到 MLP Head 进行动作分类。MLP Head 是由输入层、隐藏层、输出层构成,并且层与层之间是全连接的,采用了 GeLU 激活函数,以提高动作分类的准确性。

3 实 验

实验环境,本文提出的算法使用的硬件为 NVIDIA Tesla 4 GPU,使用 Pytorch 作为基础,保证实验的运行。

实验数据,本文选用了两个主流的人体动作识别数据集:HMDB51 数据集和 UCF101 数据集。HMDB-51 是一个相对较小的数据集,包括 51 种不同的动作类别,其中每一类都至少包含 101 个片段,总共 6 766 个视频。UCF-101 数据集是动作识别中具有挑战性的数据集,共有 101 个类别和 13 320 个视频,主要来源于 YouTube 视频库。以上数据集存在像素级低、光照差、背景混乱等缺点,为动作识别准确性带来了挑战。

实验细节,ConvTransformer 由融合通道-空间注意力机制的 CNN 结构和时间 Transformer 构成。在本文的实验里,输入部分是在整个视频中均匀地采样帧,将每个帧的较小尺寸调整为一个值∈[256,320],并从中间位置对同一视频的所有帧进行大小为 224×224 的裁剪;经过反复的实验对比发现,SAM 模块中将 2D 卷积层的卷积核设置为 7 可以获得最优性能;对于时间 Transformer,本文使用 VIT 系列模型中较小的版本,包含 6 个 MSA 层和 8 个自注意力头。

3.1 实验结果

本文将 ConvTransformer 与主流的动作识别方法进行了比较,从表 1 可以看出,Two-stream[10] 只可以捕获两个相邻帧之间的动作信息,对于依赖长时间互动的动作信息有局限性;TSN[11] 将视频进行分段处理,利用稀疏采样的方案,将长视频分割为片段进行建模,与 Two-stream 相比,显著提高了 2D CNN 基线。STAM[15] 通过将空间和时间分离,利用注意力机制进行特征强化,与 TSN 相比,性能得到了提升。但是不同帧之间的时间注意力权重差异太小,使得在建模长时间信息效果不明显。ConvTransformer

通过将 2D CNN 与时间 Transformer 相结合，在 HMDB-51 数据集和 UCF-101 数据集上，分别达到了 69.4% 和 95.5%的动作识别准确率并且优于大部分主流方法，这表明了在人体动作识别研究中引入时间 Transformer 模型建模长时间动作信息的重要性，并验证了 ConvTransformer 的有效性。

表 1　本文方法与其他方法比较

方法		HMDB-51	UCF-101
2D CNN	Two-stream[10]	59.4%	88.0%
	STCNN[13]	62.2%	90.5%
	LTC[14]	64.8%	91.7%
	TSN[11]	68.5%	94.0%
	STAM[15]	69.1%	94.3%
CNN+Transformer	ConvTransformer	69.4%	95.5%

如图 9 所示，本文选取部分视频帧可视化了以 TSN 为代表的 2D CNN 算法和 ConvTransformer“逆时针画圈”的可视化特征。可以观察到，TSN 专注于识别物体(手)而不是推理动作。与 TSN 相比，本文提出的 ConvTransformer 通过表示覆盖手画圈的特征图更好地表征动作，特别是对于图 9 可视化效果图的的第 2、3、4 列。

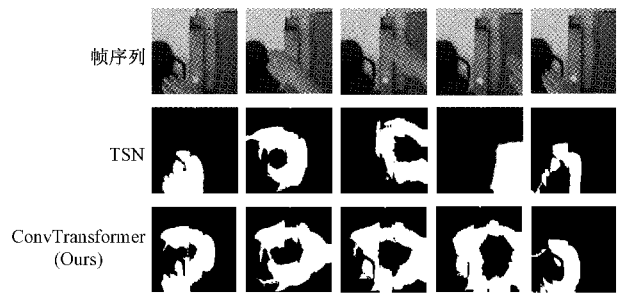

图 9　手画圈可视化效果图

3.2　消融实验

表 2 对比了加入不同模块融合的实验结果，从结果中可以看出，将 Transformer 架构引入人体动作识别领域可以显著提高动作识别的准确度，充分证明时间 Transformer 具有长期依赖性的特点，使动作识别准确率达到 95.5%。

表 2　不同模块融合在 UCF-101 数据集上的准确度

方法	准确度/%
Conv	86.1
Conv+SE	87.5
Conv+CAM	88.2
Conv+SAM	88.6
Conv+CAM+SAM	90.3
Conv+CAM+SAM+Transformer	95.5

同时，改进的 CAM 模块中同时使用平均池化和最大池化对于提取更加全面的特征也是有意义的，比 Conv+SE 在识别准确度上提升了 0.5%，也验证了本模块设计的有效性。Conv+CAM 和 Conv+SAM 分别达到了 88.2%和 88.6%的动作识别准确度，二者性能均有提升，且提升效果差距较小，充分表明通道注意力和空间注意力在提取空间特征时的贡献度相似。

在表 3 中，本文比较了作为 ConvTransformer 输入的不同序列长度。分别对 8、16 和 32 帧进行采样并比较结果。如表 3 实验结果所示，明显的趋势是随着输入序列长度的增加，精确度也随之增加。对于从 8 帧增加到 16 帧，可以看到精度提高了 1.9%；将输入帧数从 16 增长到 32，会增加 0.7%，由此可见，使用更多的帧也能够提高模型的性能。

表 3　采样不同帧数在 UCF-101 数据集上对比结果

帧数	准确度/%
8	93.6
16	95.5
32	96.2

4　结　　论

本文通过融合 2D CNN 和 Transformer 的高效视频人体动作识别模型，利用通道-空间注意力机制的 2D CNN 架构来提取帧内的空间特征；受自然语言处理领域的启发，将 Transformer 架构引入到人体动作识别研究中，利用 Transformer 捕获长期依赖性的特性，来提取不同帧之间复杂的时间信息。该算法在两个公开数据集上都得到了较好的实验结果，证明了时间 Transformer 可以很好的模拟长时间信息的动作，提高了识别精度。同时，经过消融实验对比发现，识别精度随着视频帧数的增加而增加。但是，随着帧数的增加，参数量也大幅度增加，为计算造成了负担，这也将是本文下一步的研究重点。

参考文献

[1] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset [J]. ArXiv Preprint, 2017, ArXiv: 1705.06950.

[2] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A large video database for human motion recognition [C]. 2011 International Conference on Computer Vision, IEEE, 2011: 2556-2563.

[3] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. ArXiv Preprint, 2012, ArXiv: 1212.0402.

[4] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, 2018: 7132-7141.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[6] LIU Z, ZHANG Z, CAO Y, et al. Group-free 3D object detection via transformers[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 2949-2958.

[7] ZHANG Z, SUN B, ZHANG W. Pyramid medical transformer for medical image segmentation[J]. ArXiv Preprint, 2021, ArXiv: 2104.14702.

[8] HU J, CAO L, LU Y, et al. ISTR: End-to-end instance segmentation with transformers[J]. ArXiv Preprint, 2021, ArXiv: 2105.00637.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv Preprint, 2020, ArXiv: 2010.11929.

[10] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, DOI: 10.1002/14651858.CD001941.pub3.

[11] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]. European Conference on Computer Vision. Springer, Cham, 2016: 20-36.

[12] ZHAO S, LIU Y, HAN Y, et al. Pooling the convolutional layers in deep convnets for video action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(8): 1839-1849.

[13] 于华,智敏.基于改进 CNN 框架的人体动作识别[J].计算机工程与设计,2019,40(7):2071-2075.

[14] VAROL G, LAPTEV I, SCHMID C. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1510-1517.

[15] LIN J, GAN C, HAN S. Tsm: Temporal shift module for efficient video understanding[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[16] LIU S, MA X, WU H, et al. An end to end framework with adaptive spatio-temporal attention module for human action recognition[J]. IEEE Access, 2020, DOI: 10.1109/ACCESS.2020.2979549.

[17] 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, 2017, DOI: 10.48550/arXiv.1611.06067.

[18] LIU J, WANG G, DUAN L Y, et al. Skeleton-based human action recognition with global context-aware attention LSTM networks[J]. IEEE Transactions on Image Processing, 2017, 27(4): 1586-1599.

[19] YEUNG S, RUSSAKOVSKY O, JIN N, et al. Every moment counts: Dense detailed labeling of actions in complex videos[J]. International Journal of Computer Vision, 2018, 126(2): 375-389.

[20] LI H, CHEN J, HU R, et al. Action recognition using visual attention with reinforcement learning[C]. International Conference on Multimedia Modeling. Springer, Cham, 2019: 365-376.

[21] TANG Y, TIAN Y, LU J, et al. Deep progressive reinforcement learning for skeleton-based action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5323-5332.

[22] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]. International Conference on Machine Learning. PMLR, 2021.

[23] NEIMARK D, BAR O, ZOHAR M, et al. Video transformer network[J]. ArXiv Preprint, 2021, ArXiv: 2102.00719.

[24] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding? [C]. ICML, 2021, 2(3), DOI: 10.48550/arXiv.2102.05095.

[25] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: A video vision transformer[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[26] FAN H, XIONG B, MANGALAM K, et al. Multiscale vision transformers[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[27] CORDONNIER J B, LOUKAS A, JAGGI M. On the relationship between self-attention and convolutional layers[J]. ArXiv Preprint, 2019, ArXiv: 1911.03584.

作者简介

朱相华,硕士研究生,主要研究方向为深度学习、视频图像处理。

E-mail: 335379131@qq.com

智敏(通信作者),博士,教授,主要研究方向为人工智能、深度学习、视频图像处理。

E-mail: cieczm@imnu.edu.cn

殷雁君,硕士,教授,主要研究方向为深度学习、图像处理。

E-mail: cieyyj@imnu.edu.cn