DOI: 10.13382/j.jemi.B2306948

深度嵌套注意力下的 SlowFast 信息融合动作识别网络*

张起尧 桑海峰

(沈阳工业大学信息科学与工程学院 沈阳 110870)

摘 要:视频动作识别在视频监控、自动驾驶等多个领域都有着广泛的应用。SlowFast 网络是视频动作识别领域经常使用的网络。目前 SlowFast 相关网络中使用注意力进行相关信息增强,注意力机制与网络的结合方式是将注意力机制嵌套到网络的各个卷积块之间,如果将注意力机制深层嵌套到卷积块的具体卷积层中,SlowFast 网络的信息提取能力将更进一步。首先提出了一种深度嵌套注意力机制,该深度嵌套机制内部包含一种可以提取时空与通道信息的注意力 SCTM,使 SlowFast 网络的 3 种信息提取能力得到了进一步加强。此外,目前多流网络融合的信息并没有充分的交互与处理。提出了一种基于交叉注意力与 ConvLSTM 的多流时空信息融合网络,使多流网络中每个流的信息充分交互。改进后的 SlowFast 网络在 UCF101 数据集上的 Top-1 准确率已达到 98.5%,在 HMDB51 数据集中的准确率达到了 80.1%。均优于目前已有的模型,比原始 SlowFast 网络提高了 2.64%,且鉴于上述数据,深度嵌套注意力的 SlowFast 时空信息融合网络在信息提取与融合方面具有优越性能。

关键词: 视频动作识别;SlowFast;注意力深层嵌套;信息融合网络;时空通道注意力

中图分类号: TP391;TN98 **文献标识码**: A **国家标准学科分类代码**: 520.20

SlowFast information fusion action recognition network based on deeply nested attention mechanism

Zhang Qiyao Sang Haifeng

(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: Video action recognition has been widely used in many fields such as video surveillance and automatic driving. SlowFast network is often used in the field of video action recognition. At present, attention is used to enhance relevant information in SlowFast correlation network. The combination of attention mechanism and network is to embed the attention mechanism among various convolutional blocks of the network. If the attention mechanism is deeply embedded into the specific convolutional layer of the convolutional block, the information extraction capability of the SlowFast network will be further enhanced. Firstly, a deep nested attention mechanism is proposed, which contains an attention SCTM that can extract space-time and channel information, and further strengthens the three information extraction capabilities of SlowFast network. In addition, the current multi-stream network fusion information is not fully interactive and processed. A multi-stream spatio-temporal information fusion network based on cross-attention and ConvLSTM is proposed to make the information of each stream in the multi-stream network fully interact. The improved SlowFast network has achieved 98.5% Top-1 accuracy on UCF101 and 80.1% accuracy on HMDB51. Compared with the original SlowFast network, the SlowFast spatiotemporal information fusion network with deeply nested attention has superior performance in information extraction and fusion.

Keywords: video action recognition; SlowFast; deep nesting of attention; information fusion network; spatial channel temporal attention

收稿日期: 2023-10-11 Received Date: 2023-10-11

*基金项目:国家自然科学基金(62173078)、辽宁省自然科学基金(2022-MS-268)项目资助

0　引　言

视频动作识别在视频监控、自动驾驶等多个领域都有着广泛的应用[1]。目前使用视频进行动作识别已经得到广泛应用并取得了显著的成果。精确提取视频中的时空信息并有效利用这些信息对于动作识别至关重要。

注意力机制能帮助网络更加准确地关注重要的时空特征。通过引入注意力机制,网络可以自动学习不同时空特征的重要性权重,提高对复杂动作的识别准确性。

双流网络[2]的出现确实对于视频中的时空信息的提取能力提供了显著的提升。时序流主要关注视频序列中的动态变化,而空间流则主要关注视频帧内的静态视觉特征,如纹理和形状。并通过简单的融合策略,双流网络能够更好地融合时序和空间信息,提高对视频中时空关系的建模能力。

SlowFast 网络[3]继承了双流网络与空间流与时间流的概念,通过输入不同帧率的视频区分为快慢网络,慢网络输入的是低帧率的视频帧序列。虽然 SlowFast 网络常用于视频动作识别,但是在时空信息的提取能力方面与快慢网络的信息交互与时空信息的增强方面依旧存在改进的地方,对于 SlowFast 网络的改进主要有两个方向:引入注意力机制提升时空信息提取能力[4-10],将注意力机制简单嵌套到网络各个卷积块之间增强卷积块提取相关信息的能力,但是这些方法均没有考虑注意力嵌套位置的有效性;增强快慢网络提取的时空信息融合交互能力[11-15],通过使用卷积、注意力机制或者 LSTM 等操作增强时序信息,但是这些方法均没有充分交互多流特征。针对以上问题,本文的主要贡献如下:

1)本文提出一种深度嵌套注意力机制,使得 SlowFast 网络的时空与通道信息提取能力得到了进一步加强。

2)本文提出一种基于交叉注意力与 ConvLSTM 的多流时空信息融合网络,可以有效的使 SlowFast 网络中每个流的信息充分交互,并且使融合后的信息时空特征得到进一步增强。

1　深度嵌套注意力机制下的 SlowFast 时空信息融合网络

基于视频进行动作识别一个难点是时空信息提取的准确性,本文提出的深度嵌套注意力机制下的 SlowFast 时空信息融合网络(deep nest SlowFast spatio-temporal fusion net,DSFNet)结构如图 1 所示,本文使用深层嵌套注意力机制可以更好地捕捉图像中的复杂的时空特征。对于双流网络信息的交互与处理,本文使用的交叉注意力可以有效的使快慢网络信息进行交互。相比于常见的使用注意力模块等操作增强融合后信息的时空特征,本文选择使用 ConvLSTM 模块[16],进一步增强时空特征。下面介绍各个部分的细节。

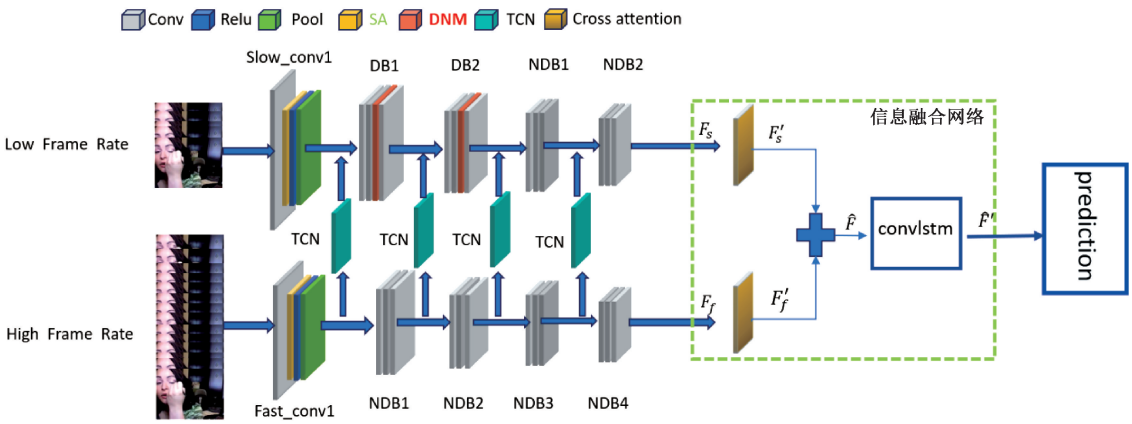


图 1　深度嵌套注意力机制下的 SlowFast 时空信息融合网络

Fig. 1　SlowFast spatio-temporal information fusion network with deeply nested attention mechanism

1.1　深度嵌套注意力机制

深度嵌套注意力机制(deep nest module,DNM)主要包含注意力机制的深层嵌套以及时空通道注意力 SCTM。将 SlowFast 网络与深层嵌套注意力机制结合,可以使一些复杂场景下,细微动作的识别准确率得到提升。

1)注意力机制的深层嵌套

注意力机制通常都是浅层地嵌套到卷积神经网络的各个卷积块之间。有些注意力机制放在网络的开端,以便在低感受野时网络更加关注相关特征。有些注意力机制放在各个模块中,以便网络能够很好地提取相关特征。还有些注意力机制放在整个网络的最后部分,以便对提取到的相关特征进行进一步的增强。显然这种方法会使

网络整体提取信息的能力得到进一步的提升,但是这种方法忽略了网络模块中各个基本单元的重要性,忽略了特征是由各个卷积进行提取,如果将注意力机制深入嵌套到卷积块中相对重要的卷积之后,就可以进一步提升该卷积提取特征的权重,之后与残差结构相连接就可以提升原先卷积模块的时空信息与通道信息的提取能力。具体的嵌套结构如图 2 所示。

卷积神经网络中每个网络层输出的特征图中的单个元素映射回原始输入特征中的区域大小,网络层越深,其输出特征的元素对应感受野越大,所以在网络深层嵌套注意力机制时可能由于浅层网络识别到的感受野并不精确导致复杂场景下动作识别的准确率较低。另外,卷积块中不同的卷积核有着不同的实际作用,如果在处理空间信息的卷积核后面加入注意力机制,就可以相当于增强了卷积块中的空间卷积能力。所以 SlowFast 网络慢网络的两个 DB 块中深度嵌套注意力机制,可以在感受野变化较少时可以感受到更为关键的空间信息。

在 SlowFast 网络中,不同的卷积模块包含不同的卷积层,每个卷积层都有各自的作用。其中慢网络中的退化块中包含两个(1 * 1 * 1)的卷积,这两个卷积的主要作用是对特征维度的变换,对信号进行简单的处理。除此之外还包含一个(1 * 3 * 3)的卷积,该卷积可以帮助网络更好地理解图像的结构和空间关系。所以对于慢网络而言退化块的(1 * 3 * 3)卷积就是相对重要的卷积,如果将注意力机制深层嵌套到在这个(1 * 3 * 3)卷积之后,相当与对(1 * 3 * 3)卷积提取的空间信号进行进一步的注意,提升该卷积的提取信息的准确率,提取后的信号又经过(1 * 1 * 1)卷积与残差链接,整个卷积模块可以更深入地学习到数据中的细节和特征,提高网络的表达能力。

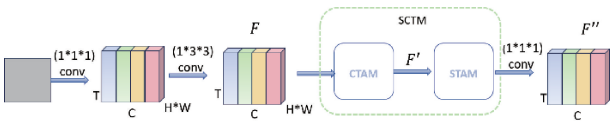


图 2　深层嵌套结构

Fig. 2　Deep nested structure

2)时空通道注意力 SCTM

SlowFast 网络中的慢网络包含一些时序信息。如果将时序信息与空间信息相结合,以时序信息辅助空间信息和通道信息,可以提高空间信息与通道信息提取的准确度。因此,本文参考 CBAM 的串联结构,将时序通道注意力与时空注意力串联使用,以提高视频动作识别的准确性。

混合注意力 CBAM 是通道注意力与空间注意力串联的结构,CBAM 引入全局平均池化与全局最大池化,可以同时提取局部特征与整体特征,但是 CBAM 并不能处理时序信息,所以本文基于 CBAM 提出一种适合 SlowFast 网络的时空通道注意力机制 SCTM。SCTM 主要分为用于取时间通道信息的 CTAM(channel temporal attention module)与用于取时空信息 STAM(spatial temporal attention module)。时间通道注意力 CTAM 具体结构如图 3 所示,通道时间注意力 STAM 具体结构如图 4 所示。

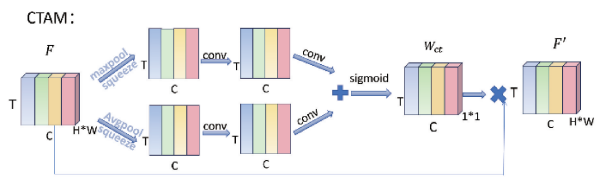


图 3　CTAM 时间通道注意力

Fig. 3　CTAM temporal channel attention

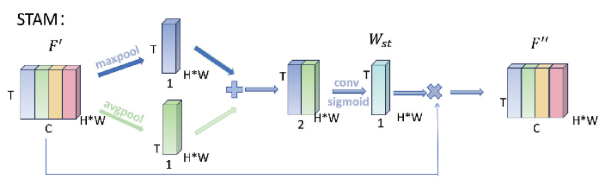


图 4　STAM 时间空间注意力

Fig. 4　STAM spatial temporal attention

本文在设计注意力时保留了 CBAM 使用两种池化分别提取不同特征的设计,在设计 CTAM 时参考 ECANet 通过使用卷积进行通道注意的方式,而不是使用原始 CBAM 通过全连接层进行通道信息的相互注意。这样操作更加考虑局部特征,可以有效的提取通道特征,除此之外因为需要考虑到时序信息对于通道信息的影响,所以在设计注意力机制的时候选择使用 3D 卷积,这样不仅可以处理通道信息,同时处理时序信息。对于输入的特征图 $F \in R^{(B,C,T,H,W)}$,其中 B 为 batch size,C 为通道数,T 为输入帧数,H,W 为输入图像的长和宽。使用 CTAM 计算通道时间注意力权重 $W_{ct}(F) \in R^{(1,C,T,1,1)}$ 的公式如式(1)所示,其中 $F^{ct}_{Max3D} = MaxPool3D(F)$,$W_0$ 与 W_1 分别表示两次卷积操作:

$$W_{ct} = \sigma(conv_{3\times3}(conv_{3\times3}(MaxPool3D(F))) + conv_{3\times3}(conv_{3\times3}(AvgP)ool3D(F)))) = \sigma(W_1(W_0(F^{ct}_{Max3D})) + W_1(W_0(F^{ct}_{Avg3D}))) \tag{1}$$

最后,通道注意力的输出权重 W_{ct} 与输入特征映射 F 按元素相乘,生成最终的通道特征映射,记为 $F' \in R^{(B,C,T,H,W)}$ 计算公式如式(2)所示。

$$F' = W_{ct} \otimes F \tag{2}$$

CTAM 模块输出的特征图 F' 作为 STAM 模块的输入特征图。首先,对特征映射 F' 进行三维全局最大池化和全局平均池化,由此产生的两个特征:AvgF' 和 MaxF',沿着通道维度拼接。然后,通过(7 * 7 * 7)卷积运算降低通道维数,再通过 sigmoid 激活函数得到空间时间注意模块的输出特征映射 W_{st}。对于输入的特征图 F' 使用 STAM 计算空间时间注意力权重 $W_{st}(F) \in R^{(1,1,T,H,W)}$ 的公式如式(3)所示:

$$W_{st} = \sigma(conv_{3\times3}(cat[MaxPool3D(F'), AvgPool3D(F')])) = \sigma(conv_{3\times3}(cat[F'^{st}_{Avg3D}, F'^{st}_{Max3D}])) \tag{3}$$

最后,通道注意力的输出 W_{st} 与输入特征映射 F' 按元素相乘,生成最终的通道特征映射,记为 $F'' \in R^{(B,C,T,H,W)}$,公式如式(4)所示。

$$F'' = W_{st} \otimes F' \tag{4}$$

1.2 交叉注意信息融合网络

SlowFast 网络的双流信息通过拼接的方式融合,各个信息之间没有交互,所以信息包含的内容较为片面,经过简单的处理就用于动作识别,会导致一些复杂场景识别准确率低,另外融合后的信息包含了大量的时序信息,同样需要一个可以同时处理空间与时间信息的模块进行进一步的增强时空信息。所以本文提出基于交叉注意力与 ConvLSTM 的信息融合网络(cross attention fusion module,CFM)用于提升双流网络信息的互补能力以及时序信息提取能力进一步提升识别的准确率。

1)基于交叉注意力的双流信息交互

SlowFast 网络横向链接的目的是使用快网络提取的时序信息帮助慢网络提取空间信息,所以在双流信息交互时采用交叉注意力机制,使用慢网络的 Q_s 作为查询向量,V_s 作为值向量,快网络的 K_f 作为键向量,通过计算 Q_s 与 K_f 的相似度,得到关注局部空间信息的特征,之后在根据相似度与 V_s 的乘积求和得到慢网络新的注意力权重。具体结构如图 5 所示。

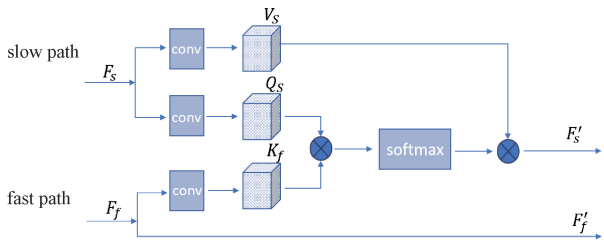


图 5 交叉注意力融合网络

Fig. 5 The network of cross-attention fusion

计算快慢网络信号 F_s 与 F_f 的交叉注意力融合的公式如式(5)所示。

$$F'_s = softmax\left(\frac{Q_s K_f^T}{\sqrt{d_k}}\right) V_s \tag{5}$$

其中,$Q_s = Conv_{3\times3}(F_s)$,$K_f = Conv_{3\times3}(F_f)$,$V_s = conv_{3\times3}(F_s)$。

为了优化双流网络的信息交互,并满足快网络时序信息对慢网络空间信息辅助设计的需求,选择快网络的时序信息作为键向量,慢网络的空间信息作为查询向量与值向量。通过计算键向量 K_f 与查询向量 Q_s 的相似度,可以得到快网络与慢网络之间相似的空间信息,从而实现对慢网络空间信息的辅助作用。

2)基于 ConvLSTM 的时空信息增强

双流网络的信息经过卷积块,注意力机制以及交叉注意力的提取之后包含了大量的时空与通道信息,需要对双流信息融合后的信息进行时空信息的增强处理,ConvLSTM 就是比较理想的增强模块,因为 ConvLSTM 不仅能够建立类似 LSTM 时序关系,而且可以拥有类似 CNN 的空间特征提取能力。在 ConvLSTM 中,门操作是基于卷积层的。这种结构允许 ConvLSTM 在处理双流网络融合后的信息 $\hat{F} \in R^{(B,C,T,H,W)}$ 时保留空间信息,并且可以自适应地学习特征和时间依赖关系。ConvLSTM 模块的结构如图 6 所示,其中蓝色虚线表示遗忘门,红色虚线表示输入门,黄色虚线表示输出门。

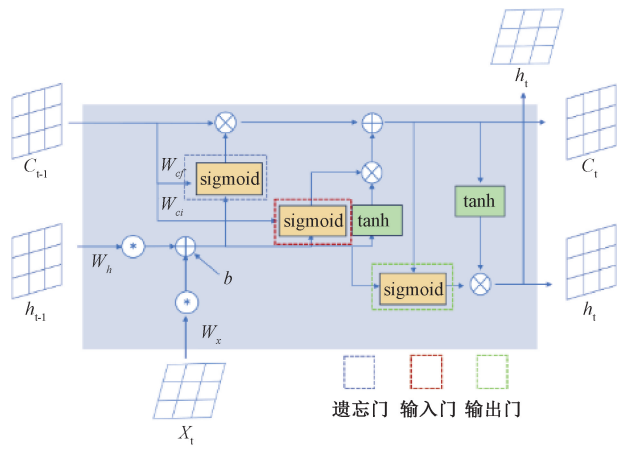


图 6 ConvLSTM 模块结构图

Fig. 6 ConvLSTM module structure

ConvLSTM 的门控单元和存储单元的方程为如下所示:

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} o c_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} o c_{t-1} + b_f) \tag{7}$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} o c_{t-1} + b_o) \tag{8}$$

$$c_t = f_t o c_{t-1} + i_t o \tanh(W_{xi} * x_t + W_{hc} * h_{t-1} + b_c) \tag{9}$$

$$h_t = o_t o \tanh(c_t) \tag{10}$$

其中,* 为卷积运算,o 为 Hadamard 积,σ 为 sigmoid

激活函数；W_x、W_h、W_c、b 分别为输入数据权值、隐藏状态数据权值、记忆状态数据权值和阈值，h_t 即为输出 $\hat{F}'$。

虽然 ConvLSTM 为现有模型，但是其在时空特征的融合方面有改进的地方，所以本文创新性的使用 ConvLSTM 与交叉注意力相结合的方式，共同组成信息融合网络，提升时空特征的融合能力。由于 SlowFast 网络结合深度嵌套注意力机制后可以很好的提取到具有重要时空价值的空间区域，交叉注意力的目的是将双流的信息进行时序方面的融合，使慢网络信息中包含的时序信息提取出来。之后在使用 ConvLSTM 对双流信息进行时空特征融合。因此使用交叉注意力与 ConvLSTM 取得了很好的融合作用。

1.3　损失函数

由于动作识别属于分类任务，所以损失函数使用的是交叉熵损失函数(Cross-EntropyLoss)，它主要刻画的是实际输出与期望输出的距离，也就是交叉熵的值越小，两个概率分布就越接近，两个概率分布越相似。若 q 为预测值，p 为真实值，则交叉熵定义如(11)所示：

$$H(p,q) = -\sum_i P(i)\log Q(i) \tag{11}$$

2　实验结果与分析

2.1　实验数据集和评价指标

本文两种数据集来评估 DSFNet 模型，UCF101 是一个现实动作视频的动作识别数据集，收集自 YouTube，提供了来自 101 个动作类别的 13 320 个视频。视频主要包括 5 大类动作。每个类别分为 25 组，每组 4～7 个短视频。

HMDB51 数据集是一个广泛用于行为识别的视频数据集，由 6 849 个视频片段组成，涵盖 51 个不同的行为类别。每个视频片段的时长约为 3～10 s。这些视频涵盖了人类的各种日常动作。

由于两种数据集都有较多的动作类别，所以 DSFNet 模型主要的评价指标为 Top-1，Top-5 准确率来衡量动作识别的准确率。Top-1 准确率是指排名第一的类别与实际结果相符的准确率，而 Top-5 准确率是指排名前五的类别包含实际结果的准确率。由于 UCF101 如果使用 Top-5 准确率作为评价指标很难看出模型识别的准确率，所以只使用 Top-1 准确率作为唯一的评价指标。

2.2　实验设置

实验环境为 Ubuntu 20.04 系统、NVIDIA 3080 显卡上进行训练和测试，参照原始 SlowFast 网络参数设置 T = 4，τ = 16，bacth size = 64，epoch = 50，learning rate = 0.001，num_workers = 4。选取在 kinetics400 数据集预训练之后得到的权重作为 SlowFast 网络的初始权重。

本文对比的模型包括：早期双流网络识别模型有 C3D、TSN、I3D、Two-Stream、TS-LSTM 等。近期的识别模型有 TSI、BQN、MSM、SVT 以及当前 state-of-the-arts (SOTA)模型 STRM 等。

2.3　实验结果与分析

1) DSFNet 与其他模型结果比较

DSFNet 与其他模型结果比较如表 1 所示。实验结果表明 DSFNet 确实可以使 SlowFast 网络对于 UCF101 数据集与 HMDB51 的识别准确率更进一步。

表 1　在 UCF101 与 HMDB51 上与其他模型比较

Table 1　Comparison with other models on UCF101 and HMDB51

模型	Top-1 准确率/%	
	UCF101	HMDB51
Two-Stream[1]	88.0	59.4
TSN[17]	94	68.5
C3D[18]	85.2	59.1
I3D[19]	93.4	66.4
TS-LSTM[20]	93.2	69.0
SIFP[21]	96.9	78.1
TSI[22]	97.2	76.9
BQN[23]	97.6	77.6
MSM[24]	93.5	66.7
SVT[25]	93.7	67.2
STRM[26]	98.1	79.3
SlowFast[3]	95.9	72.3
DSFNet	98.5	80.1

从表 1 中的数据可以看出，在 UCF101 数据集中，DSFNet 的 Top-1 (Accuracy)准确率已达到 98.5%，均优于目前已有的模型。比原始 SlowFast 网络提高了 2.64%，比目前最优的 STRM 高了 0.41%。且不需要使用 STRM 过多的计算资源，同时不需要注意特征融合导致的特征连贯性问题。在 HMDB51 数据集中准确率达到了 80.1%，比目前最优的 STRM 高了 1%.

2) 消融实验

为了验证深层嵌套注意力机制的合理性，针对 SlowFast 网络设计的 SCTM 注意力机制以及时空信息融合网络的有效性，同时为了和 SlowFast 网络进行对比，只使用 DNM 机制，保持 SlowFast 网络的融合方式，得到 DSFNet-DN。只使用将 DSFNet 的 CFM 模块，退化块与非退化块的设置不变，得到 DSFNet-CF，DSFNet-CF 中只使用 CA (cross attention)命名为 DSFNet-C，只使用 ConvLSTM，命名为 DSFNet-F。表 2 表示使用 DSFNet 使用不同模块时对比 SlowFast 网络使用 UCF101 数据集时的表现，从表 2 可以单纯使用注意力机制简单嵌套到网络(SlowFast+SCTM)中可以增加识别动作的准确率，但

是提升的幅度并不大,只有0.03%。如果将同样的注意力机制SCTM深层嵌套到SlowFast网络的DB1,DB2的Conv2之后(DSFNet-DN),动作识别的准确率提升了1.67%,这说明如果将注意力机制深层嵌套到SlowFast退化块中会极大地增强时空通道信息的提取能力。即使对SlowFast网络进行深层嵌套的注意力不是SCTM而是CBAM也同样将动作识别准确率提升了1.46%。所以将注意力机制进行深层嵌套是十分合理的。除此之外针对SlowFast网络设计的SCTM同样可以很好的提取慢网络的时空通道信息。

表2　对DSFNet不同模块的消融实验

Table 2　Ablation experiments of different modules of DSFNet

网络	模块			CFM		Top-1准确率/%	
	DNM(SCTM)	DNM(CBAM)	SCTM(浅层嵌套)	CA	ConvLSTM	UCF101	HMDB51
SlowFast	×	×	×		×	95.9	72.3
SlowFast+SCTM	×	×	√		×	96.2	75.4
DSFNet-DN	√	×	×		×	97.5	79.6
DSFNet-DN	×	√	×		×	97.3	78.9
DSFNet-CF	×	×	×		√	96.9	76.3
DSFNet-F	×	×	×	×	√	96.5	75.6
DSFNet-C	×	×	×	√	×	96.3	73.5
DSFNet	√	√	√		√	98.5	80.1

从表2可以看出DSFNet-C与DSFNet-F可以使SlowFast网络对UCF101数据集的动作识别提升0.4%,0.6%。DSFNet-C、DSFNet-F与DSFNet-CF提升幅度不高的原因同样是双流网络对于时空信息的提取能力并没有得到充分的增强,只是依靠高感受野时进行的信息交互与利用并不能极大地提升动作识别的准确率,但是同样可以起到信息增强的作用。使用完整的DSFNet不仅可以利用DNM模块极大地提升时空通道信息的提取能力,而且可以利用CFM模块极大的提升这些信息的交互与利用能力。

表3　不同模型的参数量及平均测试时长

Table 3　The number of parameters for different models and average test duration

模型	参数量/M	UCF101/s	HMDB51/s
SlowFast	33.79	248.486	168.840
DSFNet-DN	33.93	257.894	172.213
DSFNet-CF	34.03	258.671	174.735
DSFNet	34.22	263.155	177.046

从表3中可以看出通过嵌入STCM注意力机制、交叉注意力以及ConvLSTM都会使模型的参数量得到提升,提升最多的是ConvLSTM,参数量只提升了0.7%。DSFNet-DN与DSFNet-CF在UCF101数据集下平均测试时长分别提升了3.8%和4.1%,整体DSFNet的平均测试时长增加了5.9%。DSFNet-DN与DSFNet-CF在HMDB51数据集下平均测试时长分别提升了2%和3.5%,整体DSFNet的平均测试时长增加了4.9%。显然本文设计的模型是可以接受的。

深层嵌套注意力机制DNM如果使用于其中一个退化块或者深层嵌套到其他卷积层之后的实验结果如表4所示。从表4可以看出当SlowFast网络与DNM模块结合时只有同时作用于两个退化块的Conv2之后才可以发挥出DNM模块对时空通道信息的提取能力,最大程度提升整体网络的识别准确率。

表4　注意力机制深层嵌套到其他位置(UCF101)

Table 4　The attention mechanism is deeply nested in other places(UCF101)

位置	Top-1准确率			
	Conv1之前	Conv1之后	Conv2之后	Conv3之后
只用于DB1	96.8	97.2	97.4	97.1
只用于DB2	97.1	97.5	97.7	97.5
同时用于DB1,DB2	97	97.5	98.1	97.9

表5表示的是在不使用DNM模块的前提下,对于融合之后的信息使用LSTM[27],bi-LSTM[28-29],TCN[30],ConvLSTM 4种常见的时序信息提取的方式对比结果。

表5　4种时序信息提取模块对比(UCF101)

Table 5　Comparison of four temporal information extraction modules(UCF101)

模块	Top-1准确率/%
ConvLSTM	96.5
LSTM	95.1
Bi-LSTM	95.8
TCN	95.4

所以虽然LSTM,Bi-LSTM,TCN 3种模块在时序信息的提取能力方面很优秀但是对于空间信息的处理方面显得不如ConvLSTM,所以ConvLSTM是最适合SlowFast网络进行融合后信息处理的模块。

3　结　论

本文针对目前动作识别领域中注意力机制使用仅限于浅层嵌套的问题,提出了一种基于深度嵌套注意力机制的 SlowFast 时空信息融合网络。通过引入深度嵌套注意力机制,进一步加强了 SlowFast 网络的时空与通道信息提取能力。同时,为了解决目前多流网络中信息融合过于单一、融合信息缺乏充分交互的问题,提出了一种基于交叉注意力与 ConvLSTM 的多流时空信息融合网络,使得多流网络中每个流的信息能够充分交互。本文提出的网络在 UCF101 数据集上动作识别准确率提升了 2.71%,在 HMDB51 数据集中识别准确率提升了 9.5%。为动作识别领域提供了新的研究视角和方法。

参考文献

[1] 朱相华,智敏,殷雁君. 基于 2D CNN 和 Transformer 的人体动作识别[J]. 电子测量技术,2022,45(15):123-129.
ZHU X H,ZHI M,YIN Y J. Human action recognition based on 2D CNN and transformer [J] Electronic Measurement Technology,2022,45(15):123-129.

[2] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 27.

[3] CHRISTOPH F, HAOQI F, JITENDRA M, et al. SlowFast networks for video recognition [C]. IEEE International Conference on Computer Vision, 2019, 2019(1): 6201-6210.

[4] SUN N, LENG L, LIU J, et al. Multi-stream SlowFast graph convolutional networks for skeleton-based action recognition[J]. Image and Vision Computing, 2021, 109: 104141.

[5] ZHANG S, LIU H, SUN C, et al. MSTA-SlowFast: A student behavior detector for classroom environments[J]. Sensors, 2023, 23(11): 5205.

[6] DIAO X, XU Y. A SlowFast-Based violence recognition method [C]. 2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT). IEEE, 2022: 1-6.

[7] ZHANG Y, IBRAYIM M, HAMDULLA A. Research on cow behavior recognition based on improved SlowFast with 3DCBAM[C]. 2023 5th International Conference on Communications, Information System and Computer Engineering (CISCE). IEEE, 2023: 470-475.

[8] WEI D, TIAN Y, WEI L, et al. Efficient dual attention SlowFast networks for video action recognition [J]. Computer Vision and Image Understanding, 2022, 222: 103484.

[9] LI S, WANG Z, LIU Y, et al. FSformer: Fast-Slow Transformer for video action recognition[J]. Image and Vision Computing, 2023: 104740.

[10] XU K, QIN Z, WANG G, et al. Multi-focus image fusion using fully convolutional two-stream network for visual sensors [J]. KSII Transactions on Internet and Information Systems (TIIS), 2018, 12(5): 2253-2272.

[11] 袁野,黄丽清,叶锋,等. 基于集成学习双流神经网络的实时面部篡改视频检测模型[J]. 计算机工程与科学,2023,45(3):470-477.
YUAN Y,HUANG L Q,YE F,et al. A real-time facial manipulation video detection model based on ensemble learning dual-stream neural network [J]. Computer Engineering& Science,2023,45(3):470-477.

[12] SAMIEI M, CLARK J J. Predicting visual attention and distraction during visual search using convolutional neural networks[J]. arXiv preprint arXiv:2210.15093, 2022.

[13] CHEN K, XIE T, MA L, et al. A two-stream graph convolutional network based on brain connectivity for anesthetized states analysis [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2022, 30: 2077-2087.

[14] YU Y, LIU F. A two-stream deep fusion framework for high-resolution aerial scene classification [J]. Computational Intelligence and Neuroscience, 2018, 2018.

[15] 周璇,易剑平. 基于时间上下文模块的人体动作识别方法[J]. 国外电子测量技术,2022,41(10):72-79.
ZHOU X, YI J P. Human action recognition method based on temporal context network [J]. Foreign Electronic Measurement Technology, 2022, 41(10): 72-79.

[16] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting [J]. Advances in Neural Information Processing Systems, 2015, 28.

[17] 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, 2016: 20-36.

[18] DU T, LUBOMIR B, ROB F, et al. Learning spatiotemporal features with 3D convolutional networks[C]. IEEE International Conference on Computer Vision, 2015, 2015(1): 4489-4497.

[19] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.

[20] 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.

[21] LI J, WEI P, ZHANG Y, et al. A slow-i-fast-p architecture for compressed video action recognition[C]. Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, ACM, 2020: 2039-2047.

[22] SU H, FENG J, WANG D, et al. TSI: Temporal saliency integration for video action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana, USA, IEEE, 2022. 1-10.

[23] HUANG G, BORS A G. Busy-quiet video disentangling for video classification[C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, IEEE, 2022,1341-1350.

[24] ZONG M, WANG R, CHEN X, et al. Action saliency based multi-stream multiplier ResNets for action recognition[J]. Image and Vision Computing, 2021, 107: 104-108.

[25] RANASINGHE K, NASEER M, KHAN S, et al. Self-supervised video transformer[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana, USA, IEEE, 2022, 1-13.

[26] THATIPELLI A, NARAYAN S, KHAN S, et al. Spatio-temporal relation modeling for few-shot action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana, USA, IEEE, 2022. 1-15.

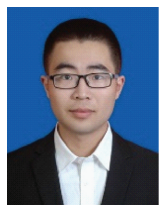
[27] 张松,李江涛,别东洋,等. 一种基于单通道 sEMG 分解与 LSTM 神经网络相结合的手势识别方法[J]. 仪器仪表学报, 2021,42(4):228-235.
ZHANG S, LI J T, BIE D Y. Gesture recognition by single-channel sEMG decomposition and LSTMnetwork[J]. Chinese Journal of Scientific Instrument, 2021,42(4): 228-235.

[28] 余金锁,卢先领. 基于分割注意力的特征融合 CNN-Bi-LSTM 人体行为识别算法[J]. 电子测量与仪器学报, 2022,36(2):89-95.
YU J S, LU X L. Human action recognition algorithm of feature fusion CNN-Bi-LSTM based on split-attention[J]. Joutnal of Electoronic Measurement and Insturmentation, 2022,36(2):89-95.

[29] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science, 2015,DOI:10.48550/arXiv.1508.01991.

[30] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint arXiv: 1803.01271, 2018.

作者简介



张起尧,2018 年于燕山大学获得学士学位,现为沈阳工业大学硕士研究生,主要研究方向为深度学习与动作识别。
E-mail: 2942084072@qq.com

Zhang Qiyao received his B. Sc. degree from Yanshan University in 2018 and is now a M. Sc. candidate at Shenyang University of Technology. His main research interests include deep learning and motion recognition.

桑海峰(通信作者),2000 年于东北师范大学获得学士学位,2003 年于东北师范大学获得硕士学位,2006 年于东北大学获得博士学位,现为沈阳工业大学教授,主要研究方向为智能视频分析、机器视觉检测与图像识别、无人驾驶之环境感知技术、深度学习技术。
E-mail: sanghaif@163.com

Sang Haifeng (Corresponding author) received his B. Sc. degree from Northeast Normal University in 2000, M. Sc. degree from Northeast Normal University in 2003 and Ph. D. degree from Northeastern University in 2006, respectively. Now he is a professor in Shenyang University of Technology. His main research interests include intelligent video analysis, machine vision detection and image recognition, autonomous driving environment perception technology, and deep learning technology.