LSCN:一种用于动作识别的长短时序关注网络

杨 珂,王敬宇,戚 琦,孙海峰,王 晶,廖建新
(北京邮电大学网络与交换国家重点实验室,北京 100876)

摘 要: 相较于图像分析,如何分析时序信息是动作识别中的一个主要问题.大多数先前的方法,如3D卷积网络、双流卷积网络,仅使用包含全局时域信息的特征作为视频的表征,忽略了局部时序特征的重要性.考虑到这样的问题,本文提出一种基于时序交互感知模块的长短时序关注网络——Long and Short Sequence Concerned Networks (LSCN),融合不同的时序信息,利用不同卷积层时序特征的交互加强对不同时序长度的动作实例的表示,兼顾长短动作实例对时序信息的需求.实验结果表明,基于3D ResNext101的LSCN在两个公共数据集(UCF101和HMDB51)上,相较于基础的网络分别有0.4%和2.9%的准确率提升.

关键词: 动作识别;时序特征;特征融合;人机交互;视频分析;深度学习
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2020)03-0503-07
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.03.012

LSCN:Concerning Long and Short Sequence Together for Action Recognition

YANG Ke,WANG Jing-yu,QI Qi,SUN Hai-feng,WANG Jing,LIAO Jian-xin
(*State Key Laboratory of Networking and Switching Technology,Beijing University of Posts and Telecommunications,Beijing* 100876,*China*)

Abstract: Compared with image analysis,how to analyze temporal information is a challenging problem in action recognition. Most of the previous methods,such as 3D CNNs (convolutional neural networks) and two-streams CNNs,only used features containing global temporal information as video representation,ignoring the importance of local temporal features. To solve this problem,we propose long and short sequence concerned networks (LSCN) based on temporal interaction perception module,which can combine different temporal information. LSCN makes use of the interactions of temporal features from different convolution layers to enhance the representation of videos and takes into account the needs of temporal information for long and short sequence actions. The results of experiments show that LSCN based on 3D ResNext101 can be generalized in two public datasets (UCF101 and HMDB51). Moreover,compared with the basic network,there are 0.4% and 2.9% accuracy improvements respectively.

Key words: action recognition; temporal feature; feature fusion; human-computer interaction; video analysis; deep learning

1 引言

近年来,由于在人机交互,监控安防等场景的应用,对视频内容的分析已经受到越来越多的关注.人类通常是视频数据中必须关注的中心,是事件的执行者,所以动作识别对视频内容的理解至关重要.从基于人工定义特征的方法[1,2]到采用深度学习的方法[3~5],动作识别的研究已经取得相当大的进展.然而,基于深度学习的视频表征的构建仍然存在一些需要解决的问题,其中由于视频中的帧随着时间的推移而演变,对时域信息的建模[5~7]至关重要.有的方法通过融合帧级特征[6~14]得到视频的时序表示.其中,基于双流卷积网络的方案[11~14]将光流模态中提取的特征作为时序表示,再与RGB模态中提取的空间特征相融合作为视频的表征.这些方法对视频中的空间信息和时间信息分开进行分析,忽略了时空信息之间的关联性,而采用3D卷积操作的卷积架构[15~23]扩展了2D卷积网络结构,同时对时域和空域进行分析,克服了这一缺点.

收稿日期:2019-03-13;修回日期:2019-05-28;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61771068,No.61671079);北京市自然科学基金(No.4182041)

3D 卷积网络的输入是由视频帧堆叠成的时空网格. 通过不断堆叠的 3D 卷积操作,网络输出的特征图尺度逐渐变小,所包含的时空信息逐渐丰富,网络最后输出的特征图代表全局的时序和外观信息,通常也被用作视频的表征. 然而,在实际的场景中存在着这样的事实,当人们观看视频时,对于某些动作实例,他们能够基于少数几帧,甚至只需一帧,就能完成对动作的判别. 例如,在图 1 射箭的视频中,我们可以在仅观察到第一帧时就能推断出动作的类别. 对于这样的动作实例,3D 卷积网络中间层的局部时序特征就可以提供足够的时序信息,而包含全局时域信息的特征就意味着时序上的冗余,对于正确地表征视频是不利的. 但是全局的时域信息对于长时序的动作实例又是必不可少的. 考虑到上述因素,本文提出一种综合考虑长短时序特征的模型——LSCN,拟合不同长度的时序特征间的交互,加强特征的时序结构,得到更优的视频表征.

图1 射箭的视频

2 相关研究

2.1 基于特征融合的动作识别

Karpathy 等人[8]使用 2D 卷积网络提取每一帧的空间特征,再融合作为视频的时空表征. RGB 输入(空间流)和光流输入(时间流)的双流网络[13],独立地计算时间和空间特征,在最后进行融合得到视频的时空表征. Feichtenhofer 等人[14,26]在双流网络的基础上,增加时间流和空间流的信息交互. TSN[12]为了表征整个视频,对整个视频稀疏地进行采样. 在 TSN 的基础上,DT-PP[11]采用时空金字塔融合帧级的特征,ECO[6]则使用 3D 卷积网络作为特征融合方案. 模拟帧和帧之间的时序关系的另一种方案是使用循环神经网络[9,10,27]. Donahue 等人[9]采用 LSTM 来整合视频帧的空间特征. 然而,RNN 在动作识别方面的表现目前落后于基于 CNN 的方法,这可能因为 RNN 的训练困难并且不能充分模拟视觉上的时序依赖[10,27].

基于特征融合的动作识别方案跨时间地融合提取自每个视频帧的空间特征以得到视频的时空表示,将时间特征和空间特征分开进行分析,忽略了时空信息间的交互. LSCN 以 3D 卷积网络作为基础模型,建立时空信息间的交互,得到的视频表征更加有效.

2.2 基于 3D 卷积网络的动作识别

Du 等人[15]引入使用 3D 卷积核的卷积网络架构,从视频的连续多帧中学习视频的时空表示. Hara 等人[16,21]沿用 3D 卷积核的使用,设计更加先进的网络结构,进一步地提升网络的表示能力. 为解决 3D 卷积网络训练困难的问题,Carreira 等人[17]提出一种新的 3D 网络参数初始化的方法,而 Xie 等人[18~20]选择对 3D 的卷积核进行时间和空间上的分解. TSM 网络[5]在时间维度上仅使用偏移操作,卷积计算在空间维度上执行,使用近似 2D 卷积的计算复杂度模拟了 3D 的卷积操作. Wang 等人[22]提出一种非局部的空间-时间模块来捕获长期的时空依赖. 为了拟合时空信息之间的关联性,Diba 等人[23]提出 STC 模块,加载到 3D 卷积网络架构当中.

3D 卷积网络架构可以对时空信息同时拟合,但是因为动作识别具有时序灵活性,即不同的动作实例需要分析的时序信息的长度不同,3D 卷积网络仅使用全局时域信息来表征视频是不足的. LSCN 在 3D 卷积网络架构的基础上,拟合不同时序特征间的交互,得到时序灵活的视频表征.

3 长短时序关注网络

在本文中,我们提出一种长短时序关注网络(LSCN),目标是通过提出一种新的模型来探索更灵活的时序结构来表征视频. 网络架构如图 2 所示,其中,3D 基础网络可以对时空信息同时进行分析,每一层卷积网络输出的特征都包含着不同大小的时域感受野. 我们首先选择两个不同层的特征作为时序交互感知模块的输入. 在时序感知模块中,对两个特征做空间上的池化操作,得到代表不同时序长度的时序特征,再进行时序融合. 时序交互感知模块融合两个代表不同长度时序的特征,拟合跨层时序特征间的交互,得到的融合特征包含更丰富的时序信息. 为了进一步丰富视频表征的时序信息,LSCN 选择多个卷积层的特征,通过多个时序交互感知模块分别分析不同时序特征之间的交互,再将所有的融合特征拼接到一起,作为视频的表征,然后馈送到动作分类器中.

3.1 时序交互感知模块

时序交互感知模块旨在拟合时序特征间的交互,将时序信息凝练到新的特征中,以探索更灵活且鲁棒的时间结构. 本文选择分解的双线性池化作为时序交互感知模块的融合策略. 在文献[28]中,双线性池化用于分析包含模态信息和时序信息的特征,拟合模态-时序间的交互,以表征 RGB-D 动作.

3D 卷积网络中卷积层的输出特征为 $\boldsymbol{M}\in R^{c\times t\times h\times w}$,其中 c 为 channels 的数量,h,w 分别为特征图的高度和宽度,t 表示时间维度. 将 $\boldsymbol{M}$ 的空间维度作池化操作,得到时序特征 $\boldsymbol{x}\in R^{c\times t}$. 为了统一两个时序特征的时间维度,对时间维度小的特征做时间维度的上采样操作. 处理后的两个时序特征可以写作:

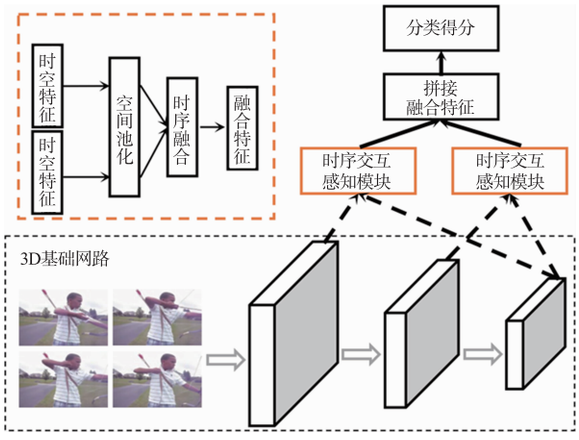


图2　LSCN的网络架构及时序交互感知模块概述

$$\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{c_1}]^{\mathrm{T}},\boldsymbol{y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{c_2}]^{\mathrm{T}}$$

c_1、c_2代表两个特征的 channels 数量. 双线性池化是可以组合两个向量空间的元素以产生第三向量空间的元素的函数,式子如下:

$$z_i=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}_i\boldsymbol{y}+\boldsymbol{b}_i \tag{1}$$

$\boldsymbol{W}_i\in R^{c_1\times c_2}$是输出向量 z_i的权重矩阵,$\boldsymbol{b}_i$是偏移值. 根据文献[29],对 $\boldsymbol{W}_i$矩阵分解得到:

$$\boldsymbol{W}_i=\boldsymbol{U}_i\boldsymbol{V}_i^{\mathrm{T}} \tag{2}$$

其中 $\boldsymbol{U}_i\in R^{c_1\times 1}$,$\boldsymbol{V}_i\in R^{c_2\times 1}$. 由式(2),式(1)可以改写为:

$$\begin{aligned}z_i&=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{U}_i\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{b}_i\\&=\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{x}\circ\boldsymbol{V}_i^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{b}_i\end{aligned} \tag{3}$$

将式(3)进行扩展可得

$$\boldsymbol{z}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}\circ\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{b} \tag{4}$$

其中 $\boldsymbol{U}\in R^{c_1\times d}$,$\boldsymbol{V}\in R^{c_2\times d}$,$\boldsymbol{b}\in R^d$ 为分解的双线性池化的参数,$\boldsymbol{z}\in R^{d\times t}$是 $\boldsymbol{x}$,$\boldsymbol{y}$ 融合后的特征. 其中 d 为两个时序特征经过时序交互感知模块融合到的新特征空间的维度.

经过式(2)的矩阵分解操作,双线性池化的参数数量减少了$(c_1\times c_2)\times d-(c_1+c_2)\times d$且仍然可微. 时序交互感知模块可以嵌入到网络中进行端到端地训练,融合时序特征,拟合时序特征间的跨层交互,得到的新的特征包含更丰富的时序信息,具有更灵活且鲁棒的时间结构.

3.2　长短时序关注网络设计

因为不同动作实例需要分析的时序长度是不相同的,即动作识别需要时序灵活的视频表征,仅使用全局的时序特征作为视频的表征,对于外观信息明显,需要分析的时序信息短的动作实例存在时序上的冗余,而局部的时序信息又不足以表征长时序的动作实例. 时序交互感知模块可以分析来自 3D 卷积架构不同层的两个时序特征间的交互,输出的特征包含两个输入特征的时序信息,有着更优秀的时间结构,可以表示更灵活的时间结构. 为了进一步加强视频表征的时间结构,LSCN 选择更多的卷积层特征,分别使用时序交互感知模块拟合两两时序信息间的跨层交互,再将得到的特征拼接到一起作为视频的表征. 相较于直接将各个不同的时序特征拼接的方案,LSCN 建立时序信息之间跨层交互,凝练特征之间的时序信息,让最后的视频表征更加紧凑.

作为基础模型的 3D 时空卷积网络最后卷积层输出的特征具有最高的语义信息,区分性最强. 为了获得更具有区分性的表征,LSCN 以自顶向下的方式将最后一层的全局时序特征向浅层的局部时序特征融合,即每个时序交互感知模块的输入都包含最后卷积层的全局时序特征和选自中间卷积层的局部时序特征,保证视频表征在有着更优秀的时间结构的同时,也能有足够的可区分性,可以有效地对动作实例进行分类.

总之,LSCN 相较于基础的 3D 卷积网络可以:(1)探索不同长度的时序特征之间的交互,凝练时域信息,提出紧凑的特征融合方案;(2)利用多个不同长度的时序特征,得到的视频表征具有更加灵活的时间结构. 在接下来,通过在 UCF101[24]和 HMDB51[25]两个数据集上的实验证明了 LSCN 的有效性.

4　实验与分析

4.1　实现细节

实验数据集　本文中的实验在两个视频动作数据集上进行评估:UCF101[22]和 HMDB51[23]. UCF101 包含 13,320 个视频,分为 101 个动作类,每个视频的平均持续时间约为 7s. HMDB51 包含 6,766 个视频,分为 51 个动作类,每个视频的平均持续时间约为 3s. 两个数据集在时间上都做了修剪,移除了非动作相关的帧,并且提供了三种不同的训练/测试划分(每种划分比例均为 70% 的视频用于训练,剩下 30% 用于测试).

网络结构　本文通过实验研究以 3D ResNeXt101[16]作为基础模型的 LSCN 的有效性. 3D ResNeXt101 的网络结构如表 1 所示. 中括号中的内容表示每个残差块的结构,中括号外的乘积表示残差块的数量. "$C=32$"意味着组卷积中组的数量为 32. 在本文中,3D ResNext101 中网络层 conv2_x 输出的特征记为 conv2_x.

训练与测试　本文采用与文献[16]相同的数据处理方案. 采样采用滑动窗口的方式,滑动窗口的大小设置为 64. 在训练阶段,窗口随机在视频上进行采样. 对于采样的视频帧,随机地进行多尺度地裁剪,裁剪的尺度包括$\left\{1,\frac{1}{2^{1/4}},\frac{1}{\sqrt{2}},\frac{1}{2^{3/4}},\frac{1}{2}\right\}$. 当尺度为 0.5 时,表示裁剪尺寸等于采样视频帧中较短边的一半长度,裁剪的位

置在帧上随机选择. 接着将视频帧以 50% 的概率进行水平翻转,然后缩放到 112 × 112 × 3 的大小. 在训练阶段,采用带动量的随机梯度下降算法来更新网络参数,动量设置为 0.9,权重衰减设置为 1e-5,批量大小设置为 32,学习率从 0.01 开始,每 20 轮将其除以 10. 选用交叉熵损失作为损失函数,训练在 60 轮的迭代后结束. 在测试阶段,采用两种评估方案以更全面地进行评估. Clip:滑动窗口在视频的第一帧开始采样,采样得到的 64 帧通过网络计算,结果即为视频的预测得分. Video:滑动窗口从视频第一帧开始滑动采样,每次滑动的步长为1,滑动至视频结束,对同一个视频的所有得分取平均值作为视频的预测得分. 两种评估方案中,对采样的每一个视频帧,在中心位置进行尺度为 1 的空间裁剪,然后缩放至 112 × 112 × 3. 本文在一块 NVIDIA TESLA M40 上使用 PyTorch 进行实验.

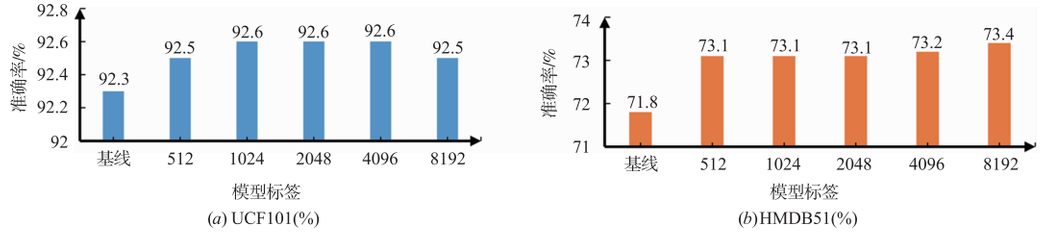
表 1 3D ResNeXt101 的网络结构

网络层	特征图的大小	ResNeXt101
conv1	64 × 56 × 56 × 64	7 × 7 × 7,64,stride 1,2,2
conv2_x	32 × 28 × 28 × 256	3 × 3 × 3,max pool,stride 2
		$\begin{Bmatrix}1\times1\times1,128\\3\times3\times3,128,C=32\\1\times1\times1,256\end{Bmatrix}\times3$
conv3_x	16 × 14 × 14 × 512	$\begin{Bmatrix}1\times1\times1,256\\3\times3\times3,256,C=32\\1\times1\times1,512\end{Bmatrix}\times4$
conv4_x	8 × 7 × 7 × 1024	$\begin{Bmatrix}1\times1\times1,512\\3\times3\times3,512,C=32\\1\times1\times1,1024\end{Bmatrix}\times23$
conv5_x	4 × 4 × 4 × 2048	$\begin{Bmatrix}1\times1\times1,1024\\3\times3\times3,1024,C=32\\1\times1\times1,2048\end{Bmatrix}\times3$
classifier	2048	global average pool
	类别数量	fc

4.2 映射空间的维度选取

时序交互感知模块将两个时序长度的特征融合到新的特征空间中去,融合后的特征 $z \in R^{d\times t}$,其中 t 表示特征的时间维度,d 表示融合后新的空间维度,为了研究新的空间维度 d 的选取对实验结果的影响并验证基于 3D ResNext101 的 LSCN 的有效性,本节对 UCF101[22] 和 HMDB51[23] 中的第一种划分进行实验对比,结果总结在图 3 中.

其中,横轴上的基线表示 3D ResNext101,512、1024、2048、4096、8192 分别表示当 d 值取对应值时的 LSCN,纵轴上的数值表示对应模型的动作识别准确率的百分比值. 柱状图下方的 UCF101(%)、HMDB51(%)表明了相应结果使用的实验数据集分别为 UCF101 split1 和 HMDB51 split1. 在本节实验中,时序特征选择 conv5_x、conv4_x 和 conv5_x、conv3_x. 由图 3 可得,LSCN 的结果要优于基线(3D ResNext101[16]),但是 d 值的选取对于 LSCN 的影响并不大. 在 UCF101 split1 上,当 d 值较小时,融合后的特征维度可能过小,区分性弱,准确率较低,随着 d 的增大,LSCN 的准确率逐渐变高,当 d 等于 1024,2048,4096 时,准确率达到最大值,而随着 d 的继续增大,准确率开始下降,可能的原因是 d 值过大,带来过多的参数,网络变得倾向过拟合. 而在 HMDB51 split1 上,整体趋势表现为 LSCN 的准确率随着 d 值的增大而增大. 可见,在不同的数据分布上,LSCN 对于 d 值的适应性稍有不同,但都是有效的. 在接下来的实验中,为了进行公平的比较,在两个数据集上 d 值都设置为 2048.

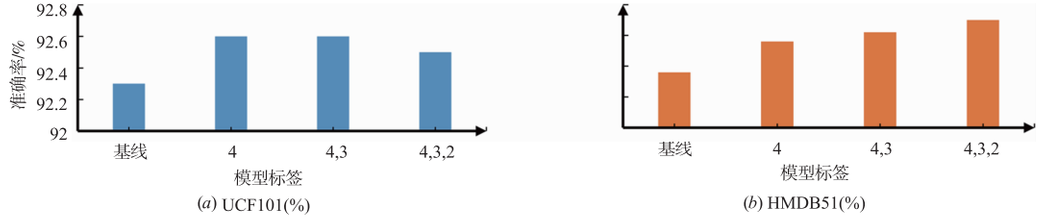


(a) UCF101(%) (b) HMDB51(%)

图3 d值对LSCN的性能影响.图示分别为UCF101 split1和HMDB51 split1上Clip的准确率

4.3 局部时序特征的选择

LSCN 选用多个局部时序特征来得到更灵活的时序结构,本节通过实验研究局部时序特征选取对 LSCN 性能的影响,选择的每一个局部时序特征都与全局时序特征 conv5_x 进行融合. 实验的结果如图 4 所示. 其中,横轴上的基线表示 3D ResNext101,4、4、3、4、3、2 分别表示选取相对应的卷积特征的 LSCN,例如,项目“4,3”意味着局部时序特征选择 conv4_x 和 conv3_x 再分别与 conv5_x 组成两个时序交互感知模块构成的 LSCN. 纵轴上的数值表示对应模型的动作识别准确率的百分比值. 柱状图下方的 UCF101(%)、HMDB51(%)表明了相应结果使用的实验数据集分别为 UCF101 split1 和 HMDB51 split1.



(*a*) UCF101(%)　　(*b*) HMDB51(%)

图4　选用不同的时序特征对LSCN的性能影响,图示分别为UCF101 split1和HMDB51 split1上Clip的准确率

由图4可得,在UCF101 Split1上,LSCN的Clip准确率相较于基线有0.2%到0.3%的提高,而在HMDB101 Split1上,则有1%到1.5%的提高.在UCF101 split1上,当局部时序特征选择conv4_x和conv3_x或者仅选择conv4_x时LSCN取得最好的结果,而在HMDB51 split1上,选择conv4_x、conv3_x和conv2_x时LSCN取得最好的结果,而且时序特征的选取对HMDB51 split1上模型性能的影响较大.在接下来的实验中,为了统一模型结构以进行公平的比较,我们选择conv4_x和conv3_x两个局部时序特征组合成两个时序交互感知模块构建LSCN.

4.4　LSCN性能分析

图5显示了基础模型和LSCN在微笑,大笑,灌篮,传球几个实例的置信度得分.由图5可知,LSCN对于空间信息相近,时序信息的分析要求更高的动作实例有着更好的表现.

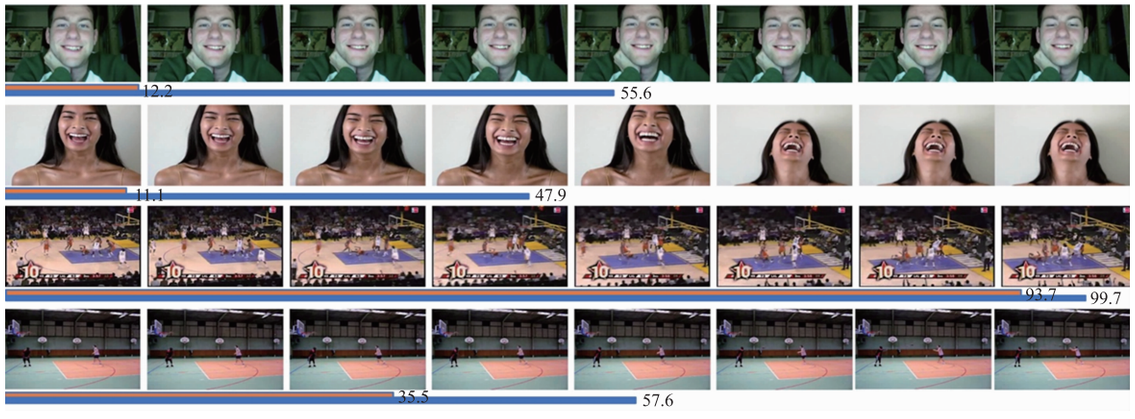


图5　从上至下的动作分别为:微笑、大笑、灌篮、传球.橙色轴线、蓝色轴线分别代表基础模型、LSCN的预测得分

图6报告了LSCN在UCF101和HMDB51的第一种划分上的训练集和测试集的损失和训练轮次的曲线.可见,LSCN可以在动作识别任务上有效地收敛.

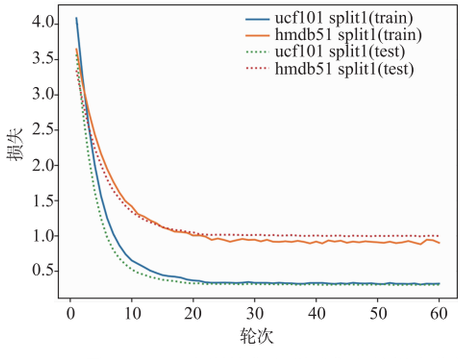


图6　LSCN在训练集和测试集上的损失曲线

表2报告了3D ResNext101和LSCN的计算量和运行时间的比较.表2中的数据表明LSCN在3D ResNext101的基础上增加了0.2%的计算量和30%的运行时间,因为时序交互感知模块的实现包含了较多的矩阵维度变换操作,虽然没有引入较多的参数和计算量,但是时间花费较多.

表2　计算量和运行时间比较

方案	计算量(百万)	运行时间(64帧)
3D ResNeXt101	38268	0.198
LSCN	38275	0.259

本文使用3D ResNeXt101[16]作为基础的网络.我们按照文献[16]复现两个数据集上的实验.在UCF101[22]上,复现的结果没有达到[16]报告的结果(93.7% vs 94.5%),然而,在HMDB51[23]上,却得到了更好的结果(70.6% vs 70.2%).本文采用实际复现的结果进行评估.表3报告了在UCF101[22]和HMDB51[23]上的所有划分中各个方案的准确率的平均值.

表3显示,LSCN在UCF101和HMDB51上相对于

3D ResNext101 分别有 0.4% 和 2.9% 的准确率提升. 而 Two-stream I3D 则在两个数据集中都取得了最好的结果,原因是 Two-stream I3D 的所有网络层都在 Kinetics 上进行了预训练而且输入的模态包含 RGB 信息和光流信息,而 LSCN 仅使用 RGB 模态并且没有在 Kinetics 上预训练全部的网络(时序交互感知模块采用随机初始化的方案),仍然获得了不错的结果. 同时还可以看出,LSCN 在 HMDB51 上的改进要优于 UCF101. 可能有以下两种原因:(1) UCF101 上的基础模型的准确率在 94% 左右,已经得到较为理想的效果,得到更高的提升是较为困难的. (2) HMDB51 中的动作类别包括细致的咀嚼、微笑,也包括较为简单的运球、骑马,不同动作实例需要的时序信息差异更大,时间结构变化更复杂,更符合 LSCN 对时间结构的设计.

表 3 UCF**101** 和 HMDB**51** 数据集上准确率比较

方案	预训练	UCF101 (%)	HMDB51 (%)
TSM(offline)[5]	Kinetics	96.0	73.2
TSN[12]	ImageNet + Kinetics	94.2	69.4
DTPP[11]	Kinetics	89.7	61.1
ECO[6]	Kinetics	94.8	72.4
C3D[15]	Sports-1M	85.8	54.9
Two-stream I3D[17]	ImageNet + Kinetics	**98.0**	**80.7**
T3D[21]	Kinetics	91.7	61.1
3D ResNeXt101[16]	Kinetics	93.7	70.6
LSCN	Kinetics	**94.1**	**73.5**

5 结束语

本文提出基于分解的双线性池化的时序交互感知模块,可以利用时序特征之间的信息交互得到更有效且紧凑的时序特征. 在此基础上,长短时序关注网络(LSCN)利用多个不同时序长度的特征,分别使用时序交互感知模块拟合两两时序特征间的跨层交互,构建具有更加灵活时间结构的视频表征. 此外,我们通过实验研究融合空间维度大小及局部时序特征的选取对 LSCN 性能的影响. 最后,基于 3D ResNeXt101 的 LSCN 在 UCF101 上将动作识别的准确率从 93.7% 提升到了 94.1%,在 HMDB51 上将动作识别的准确率从 70.6% 提升到了 73.5%,证明了 LSCN 可以得到具有更强的时间结构的视频表征.

参考文献

[1] WANG H, SCHMID C. Action recognition with improved trajectories[A]. International Conference on Computer Vision[C]. Australia: IEEE, 2013. 3551 – 3558.

[2] WANG L, Qiao Y, TANG X. MoFAP: A multi-level representation for action recognition[J]. International Journal of Computer Vision, 2016, 119(3): 254 – 271.

[3] SUN S, KUANG Z, SHENG L, et al. Optical flow guided feature: a fast and robust motion representation for video action recognition[A]. Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 1390 – 1399.

[4] LEE M, LEE S, SON S, et al. Motion feature network: Fixed motion filter for action recognition[A]. European Conference on Computer Vision (ECCV)[C]. Germany: Springer, 2018. 387 – 403.

[5] LIN J, GAN C, HAN S. Temporal Shift Module for Efficient Video Understanding [DB/OL]. arXiv: 1811.08383, 2018.

[6] ZOLFAGHARI M, SINGH K, BROX T. ECO: Efficient convolutional network for online video understanding[A]. European Conference on Computer Vision[C]. German: Springer, 2018. 713 – 730.

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

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

[9] DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[A]. Computer Vision and Pattern Recognition[C]. USA: IEEE, 2015. 2625 – 2634.

[10] LEV G, SADEH G, KLEIN B, et al. Rnn fisher vectors for action recognition and image annotation[A]. European Conference on Computer Vision[C]. German: Springer, 2016. 833 – 850.

[11] ZHU J, ZHU Z, ZOU W. End-to-end video-level representation learning for action recognition[A]. International Conference on Pattern Recognition (ICPR)[C]. China: IEEE, 2018: 645 – 650.

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

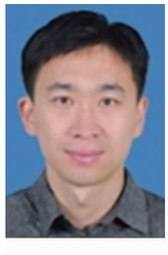
[13] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[A]. Neural Information Processing Systems[C]. Canada: NIPS, 2014. 568 – 576.

[14] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[A]. Computer Vision and Pattern Recognition

[C]. USA:IEEE,2016. 1933 – 1941.
[15] TRAN D, BOURDEV L D, FERGUS R, et al. C3D: generic features for video analysis[J]. Computer Research Repository, 2014, 2(7): 1 – 8.
[16] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet [A]. Computer Vision and Pattern Recognition [C]. USA: IEEE, 2018. 18 – 22.
[17] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[A]. Computer Vision and Pattern Recognition(CVPR)[C]. USA: IEEE, 2017. 4724 – 4733.
[18] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[A]. Computer Vision and Pattern Recognition [C]. USA: IEEE, 2018. 6450 – 6459.
[19] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[A]. International Conference on Computer Vision(ICCV)[C]. Italy: IEEE, 2017. 5534 – 5542.
[20] XIE S, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning for video understanding[J]. Computer Research Repository, 2018, 27(7): 1 – 10.
[21] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3D convNets: New Architecture and Transfer Learning for Video Classification[DB/OL]. arXiv: 1711. 08200, 2017.
[22] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[A]. Computer Vision and Pattern Recognition[C]. USA: IEEE, 2018. 7794 – 7803.
[23] DIBA A, FAYYAZ M, SHARMA V, et al. Spatio-temporal channel correlation networks for action classification [A]. European Conference on Computer Vision[C]. German: Springer, 2018. 284 – 299.
[24] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. Computer Science, 2012, 3(12): 2 – 9.
[25] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[A]. International Conference on Computer Vision (ICCV) [C]. Spain: IEEE, 2011. 2556 – 2563.
[26] FEICHTENHOFER C, PINZ A, WILDES R. Spatiotemporal residual networks for video action recognition [A]. Neural Information Processing Systems[C]. Spain: NIPS, 2016. 3468 – 3476.
[27] LI Z, GAVRILYUK K, GAVVES E, et al. VideoLSTM convolves, attends and flows for action recognition[J]. Computer Vision and Image Understanding, 2018, 166 (3): 41 – 50.
[28] HU J F, ZHENG W S, PAN J, et al. Deep bilinear learning for rgb-d action recognition[A]. European Conference on Computer Vision (ECCV) [C]. German: Springer, 2018. 335 – 351.
[29] RENDLE S. Factorization machines [A]. International Conference on Data Mining[C]. USA: IEEE, 2010. 995 – 1000.

作者简介

杨　珂　男,1995 年 9 月出生于辽宁省丹东市. 北京邮电大学网络与交换国家重点实验室硕士研究生. 研究方向为视频动作识别.
E-mail:18811582632@163. com



王敬宇(通信作者)　男,1978 年 2 月出生于吉林省长春市. 北京邮电大学网络与交换国家重点实验室教授、博士生导师. 研究方向为智能网络、人工智能、云计算、多媒体通信、多路径传输、流量工程.
E-mail:wangjingyu@bupt. edu. cn