计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(1):204−211 [doi: 10.15888/j.cnki.csa.008242]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于时移和片组注意力融合的双流行为识别网络①

肖子凡[1,2,3], 刘逸群[4], 李楚溪[5], 张　力[6], 王守岩[1,2,3], 肖　晓[2,3]

[1](复旦大学 工程与应用技术研究院 上海智能机器人工程技术研究中心, 上海 200433)
[2](计算神经科学与类脑智能教育部重点实验室 (复旦大学), 上海 200433)
[3](复旦大学 类脑智能科学与技术研究院, 上海 200433)
[4](复旦大学 计算机科学技术学院 上海市智能信息处理重点实验室, 上海 200433)
[5](复旦大学 信息科学与工程学院 微纳中心, 上海 200433)
[6](复旦大学 大数据学院, 上海 200433)
通信作者: 肖　晓, E-mail: xiaoxiao@fudan.edu.cn

摘　要: 基于深度学习的行为识别算法往往由于复杂的网络设计而难以在实际应用中达到快速、准确的识别效果. 针对以上情况, 提出一种轻量型的基于时移和片组注意力融合的端到端双流神经网络模型. 算法在 RGB 与光流分支网络中, 采用时间稀疏分组随机采样策略实现长时程建模, 利用时移模块在时间维度上置换部分通道从而结合邻帧信息来提升时序表征能力, 同时通过多路径及特征图注意融合的片组注意力模块提升网络的识别性能. 实验表明, 模型在行为识别公共数据集 UCF101 及 HMDB51 上分别达到了 95.00% 和 72.55% 的识别准确率.

关键词: 行为识别; 双流深度网络; 时移模块; 片组注意力

引用格式: 肖子凡,刘逸群,李楚溪,张力,王守岩,肖晓.基于时移和片组注意力融合的双流行为识别网络.计算机系统应用,2022,31(1):204–211. http://www.c-s-a.org.cn/1003-3254/8242.html

Two-stream Action Recognition Network Based on Temporal Shift and Split Attention

XIAO Zi-Fan[1,2,3], LIU Yi-Qun[4], LI Chu-Xi[5], ZHANG Li[6], WANG Shou-Yan[1,2,3], XIAO Xiao[2,3]

[1](Shanghai Engineering Research Center of AI & Robotics, Academy of Engineering and Technology, Fudan University, Shanghai 200433, China)
[2](Key Laboratory of Computational Neuroscience and Brain-inspired Intelligence, Ministry of Education (Fudan University), Shanghai 200433, China)
[3](Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China)
[4](Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China)
[5](Micro Nano System Center, School of Information Science and Technology, Fudan University, Shanghai 200433, China)
[6](School of Data Science, Fudan University, Shanghai 200433, China)

Abstract: The deep learning-based algorithms of action recognition are often difficult to achieve fast performance and high accuracy due to the complexity of neural networks. In view of this, we modularize the existing temporal shift and split attention module as an end-to-end trainable block which can be easily plugged into the classical two-stream action recognition pipeline. In the RGB and optical flow branch network, we adopt a random sampling strategy with sparse temporal grouping to realize long-term modeling. Furthermore, we use the Temporal Shift module to replace some channels in the time dimension so as to enhance the sequential characterization ability with information of adjacent frames. In addition, the Split Attention module integrating multi-paths and feature map attention mechanism improves the recognition performance of the network. Experiments show that our method achieves appealing performance on two

① 基金项目: 国家重点研发计划 (2019YFA0709504); 国家自然科学基金青年项目 (31900719); 上海市科技人才计划启明星项目 (19QA1401400); 上海市市级重大科技专项 (2018SHZDZX01)
收稿时间: 2021-03-20; 修改时间: 2021-04-16; 采用时间: 2021-04-20; csa 在线出版时间: 2021-12-17

© 中国科学院软件研究所 http://www.c-s-a.org.cn

public benchmark datasets including UCF101 (recognition accuracy of 95.00%) and HMDB51 (recognition accuracy of 72.55%), demonstrating its effectiveness.

Key words: action recognition; two-stream deep network; temporal shift module; split attention

计算机视觉是使用计算机及相关设备对生物视觉机制的一种模拟技术. 在图影资料剧烈增长的信息化时代, 如何智能感知和解读图影, 成为了计算机视觉领域重要的研究方向. 其中, 行为识别作为计算机视觉领域的一个重要应用分支, 已在智能监控[1]、异常行为检测[2]、人机交互[3]、视频预测[4]、医疗健康[5] 等众多领域扮演着越来越重要的角色, 具有十分广阔的应用前景.

行为识别的实现方法可分为传统的机器学习方法和深度学习方法. 传统的机器学习方法的优势在于模型简单、分类速度快, 代表性的方法有 iDT (improved dense trajectories)[6] 算法, 其使用改进的特征编码方式来表征人体运动, 但基于密集的流场运算会产生高维数据特征, 这将大大增加存储开销. 而近 10 年, 基于神经网络的深度学习方法凭借模拟人类神经元的传递原理、复杂网络的设计、参数的反向传播机制以及端到端的架构使其成为直接输出结果的任意复杂函数逼近器, 逐渐成为了视觉任务的主流方法, 并且被证实比传统机器学习方法更加强大和鲁棒[7], 深度学习算法从而也被广泛运用到行为识别任务中. 而基于神经网络的行为识别主要分为两个过程: 特征表示与动作的感知及理解.

针对视频行为识别任务, 目前的深度学习方法可分为基于 3D 卷积神经网络 (3D CNN)[8,9]、长短记忆单元 (LSTM)[10−12] 及双流神经网络 (Two-Stream CNN)[13−16] 的行为识别模型. 其中 3D CNN 可以捕获时空特征, 这意味着它可直接对视频进行特征提取, 因此具有较好的识别性能, 但其需要训练大量数据的同时也会产生较高的计算开销. 而 LSTM 虽然具有处理时序数据的先天优势, 但其容易引起梯度消失且不能很好地学习时序之外的横向信息, 比如运动特征. Two-Stream CNN 最早由 Karpathy 等人[14] 提出, 它通过扩展 CNN 局部时空信息以达到时空域上的连通性, 并通过分析额外的运动信息对 CNN 预测性能的影响, 从而选择两个输入流进行不同分辨率特征的学习, 通过融合多尺度时空信息从而提高了网络的识别精度. 与此同时, 在基于视频的输入策略上, 以往的密集采样往往带来较高的计算量且未能对长时程行为进行有效建模, 而固定间距采样的堆帧并不能保证特征信息的有效利用, 从而不能有效提高网络的泛化能力.

针对以上问题, 本文首先在整体输入上采取时间稀疏分组随机采样策略[17], 从而保证有效信息的长时程覆盖. 同时创新性地提出基于时移 (temporal shift, TS)[18] 和片组注意力 (split attention, SA)[19] 模块融合的轻量型时空双流网络模型 (TS-SA Net). 其中时移模块可以让模型在二维卷积的基础上学习到时序特征, 片组注意力机制则用于帮助网络“聚焦”有效区域, 从而产生更具分辨性的特征, 提高网络的行为识别能力. 目前我们在 UCF101、HMDB51 上分别取得了 95.00% 和 72.55% 的识别精度.

1 基于时移和片组注意力融合的时间分组双流网络 (TS-SA Net)

1.1 TS-SA Net 整体架构

行为识别任务的本质是分类 (classification) 问题, 即给定一个待识别的样本 x^q 和包含 $D^g\{x_i^g \in D^g | i = 1, 2, 3, \cdots, D^g\}$ 个样本的数据集, 算法需要依据数据集学习不同行为类别的标识特征, 从而将待识别样本与映射空间做高维距离计算, 并将其归纳入与之特征差异最小的类族中.

在本文提出的双流 TS-SA 网络中, 针对待处理视频集 $V_{\text{Data}} = \left\{V^1, V^2, \cdots, V^M\right\}$ 中, 我们将每一视频 V (采样后) 逐帧输入 TS-SA 网络, 假设帧输入为 $I_p^q \in V_{\text{Data}}, q \in \{1, 2, \cdots, M\}, p \in \{1, 2, \cdots, N\}$, 其中 N 为单个视频所含中图片序列数目. 如图 1 所示, 基于 TS-SA 网络的行为识别过程可形式化为:

$$P\left(I_1^q, I_2^q, \cdots, I_p^q\right) = H\left(G\left(F\left(I_1^q, W\right), F\left(I_2^q, W\right), \cdots, F\left(I_p^q, W\right)\right)\right) \tag{1}$$

其中, $F\left(I_p^q, W\right)$ 是参数为 W 的卷积函数, 为每一输入帧 I 产生片段级类别得分. G 为片段聚合函数, 用于整合各片段的判决分数并得到视频唯一的类别得分. 当 RGB 网络与光流 (flow) 网络各自产生视频级的预测结果后, 设计预测函数 H 来对整个视频进行动作类别的概率预

© 中国科学院软件研究所 http://www.c-s-a.org.cn

测, 本文使用 Softmax 函数.

1.2 时间稀疏分组随机采样策略

对于深度神经网络来说, 数据集及数据的采样对结果十分重要, 其往往决定了网络的学习质量与效率. 在采样策略上, 虽然减少图像序列的输入能够直接降低计算量, 但这同时造成了行为内容本身的缺失, 尤其无法对长时程行为进行完整建模. 而密集的图像序列输入虽然能保证行为特征的完全捕获, 但大量的数据带来了高额计算, 使网络缓慢笨重. 而固定间隔的图像采样方法同样存在有效信息遗失的问题.

我们意识到, 对于包含特定行为的视频来说, 相邻帧所包含的信息是高度重叠的, 这为网络非密集输入条件下保持性能的稳定提供了事实基础. 基于此, 本文在输入策略上采用了稀疏分组随机采样方法.

如图 1 所示, 本文所提出的 TS-SA Net 使用时间稀疏分组随机采样的视频帧作为输入. 具体地, 我们将视频进行等间距地稀疏分离为N个片段, 即$V^i=\{S_1,S_2,\cdots,S_N\}$, $i\in\{1,2,\cdots,M\}$. 对视频片段S_N进行随机抽样得到T_N, 则视频V^i的输入形式为$\{T_1,T_2,\cdots,T_N\}$. TS-SA 网络的识别函数如式 (2):

$$P(T_1,T_2,\cdots,T_N)=H(g(F(T_1,W),F(T_2,W),\cdots,F(T_N,W))) \tag{2}$$

具体地, N个视频片段间参数共享, 使用标准分类交叉熵作为损失函数, 如式 (3):

$$L(y,G)=-\sum_{i=1}^{C}y_i\left(G_i-\log\sum_{j=1}^{C}\exp G_j\right) \tag{3}$$

其中, C为动作类别数, y_i为属于第i类的真实标签, G_i为第i类的预测结果, 由聚合函数g得出:

$$G=g(F(T_1,W),F(T_2,W),\cdots,F(T_N,W)) \tag{4}$$

实验中, 我们试验了多种不同的聚合函数g, 发现平均融合的效果最优, 具体可见第 2.3 节. 在使用多个片段共同优化网络的过程中, 利用反向传播算法调整模型参数, 如式 (5):

$$\frac{\partial L(y,G)}{\partial W}=\frac{\partial L}{\partial G}\sum_{n=1}^{N}\frac{\partial g}{\partial F(T_n)}\frac{\partial F(T_n)}{\partial W} \tag{5}$$

具体实践中使用随机梯度下降 (SGD) 优化模型, 设置稀疏分组数的超参N为 8, 保证了参数的优化是依据结合了所有采样帧 (视频片段) 的预测结果, 利用非密集的数据输入, 从视频层构建行为识别模型.

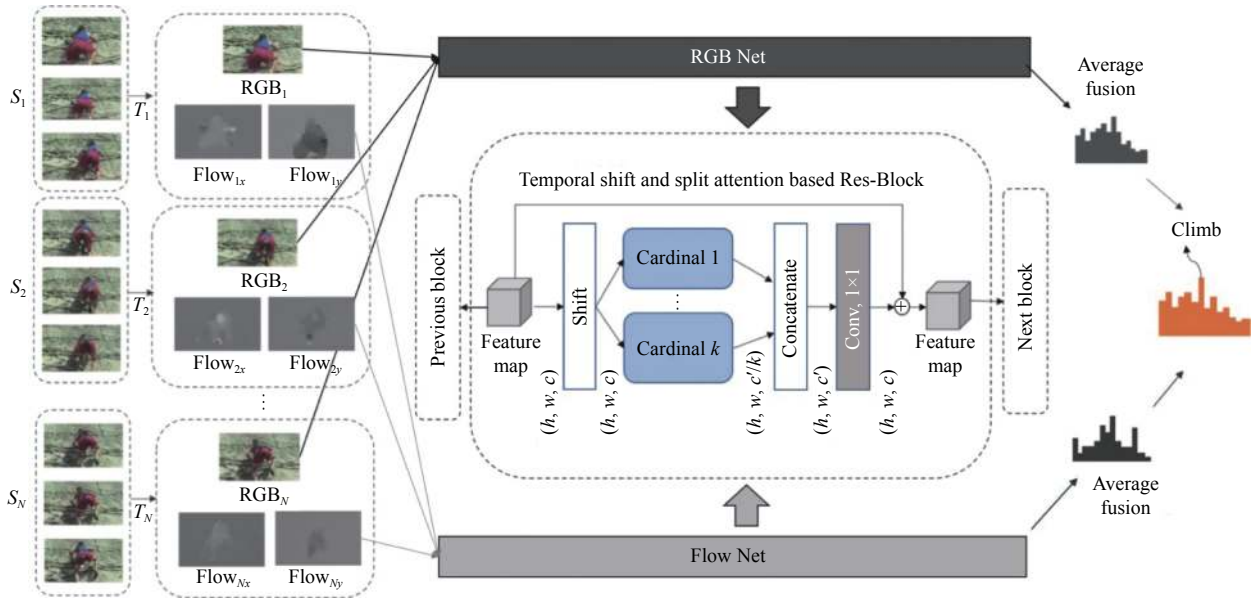

图 1 基于时移和片组注意力的双流网络 (TS-SA Net) 的结构

1.3 片组注意力模块

以 SK-Net[20] 为代表的多路径 (multi-path) 注意力启发自人脑皮质神经元根据不同的刺激可动态调节自身的感受野, 是一种通过非线性地融合来不同分支下的核尺寸对应的特征来捕获不同比例的目标对象的动态选择注意机制. SE-Net[21] 则通过重新定义通道间特征图谱的关系来实现“特征重标定”, 即对于不同通道的特征来说, 加强有效信息的权重并压缩无用信息的

© 中国科学院软件研究所 http://www.c-s-a.org.cn

参与, 它属于一种通道层级的注意力机制——自适应地调整通道特征响应. 前者在 ResNeXt[22] 的基础上用不同分支对应的不同尺寸的卷积核减少计算量而维持性能不变, 后者建立了通道层级的注意机制, 可自适应地学习不同通道间的特征关系. 在面对深度学习中大量的矢量计算时, 基于多路径和通道的注意机制都只在通道维度对权值进行了重分配, 而未考虑特征图谱内的关系响应, 所以两者的提升效果有限.

本文采用基于残差块 (residual block)[23] 的多路径与特征注意结合的注意力映射方法, 使得注意力得以跨特征图谱运作, 图 2 展示的是一个片组注意力模块.

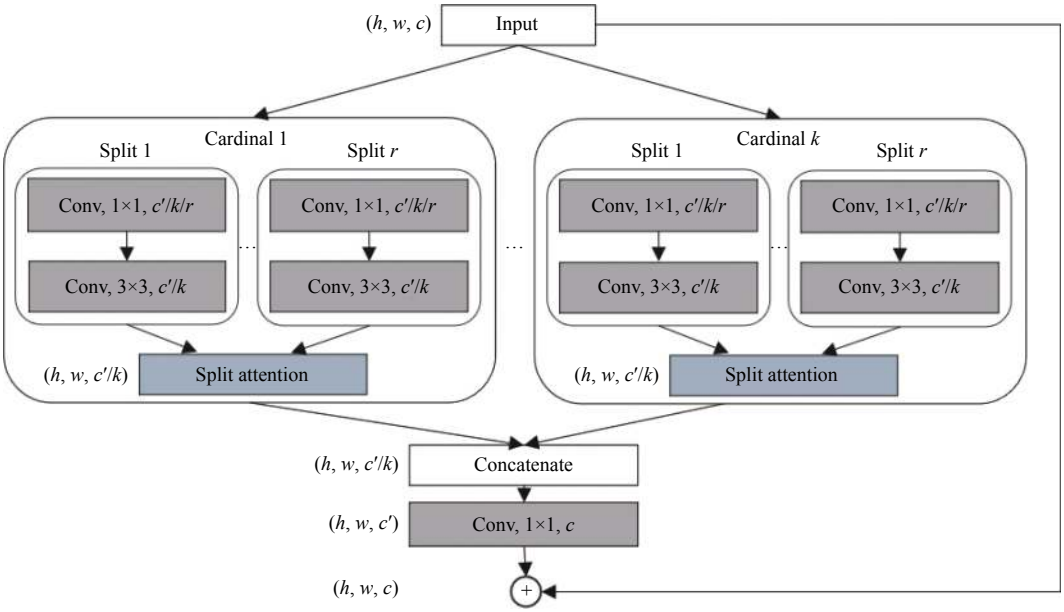

图 2 片组注意力模块

在 RGB 和光流分支网络中, 特征首先被分为几个基组 (cardinal), 每一基组再划分为若干片组 (split)(详细结构于图 2、图 3), K 和 R 分别是基组数和片组数的超参, 因此特征组的总数为$G=KR$, 实验中分别设置为$K=2$, $R=4$. 我们对不同特征图组采用不同的学习函数$\{\mathcal{F}_1,\mathcal{F}_2,\cdots,\mathcal{F}_G\}$, 则每一组的学习特征可表示为$U_i=\mathcal{F}_i(X), i\in\{1,2,\cdots,G\}$, 其中$\mathcal{F}_i$为$1\times1$卷积和$3\times3$卷积的组合, 如图 2 所示.

具体地, 每个基组的映射算法设计为多个片组的元素 (element-wise) 加和结果, 因此第k个基组的表达如式 (6):

$$\hat{U}^k=\sum_{i=R(k-1)+1}^{Rk}U_i \tag{6}$$

其中, $\hat{U}^k\in\mathbb{R}^{H\times W\times C/K}, k\in\{1,2,\cdots,K\}$, H, W和C为分块输出的特征图谱的尺寸.

在每个基组中, 首先通过跨越空间维度的全局平均池化可以收集全局上下文信息, 如图 3. 设$s^k\in\mathbb{R}^{C/K}$表示第k个基组的全局平均池化结果, s_c^k为基组中第$c(c=C/K)$个分量, s_c^k的计算公式如式 (7):

$$s_c^k=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{U}_c^k(i,j) \tag{7}$$

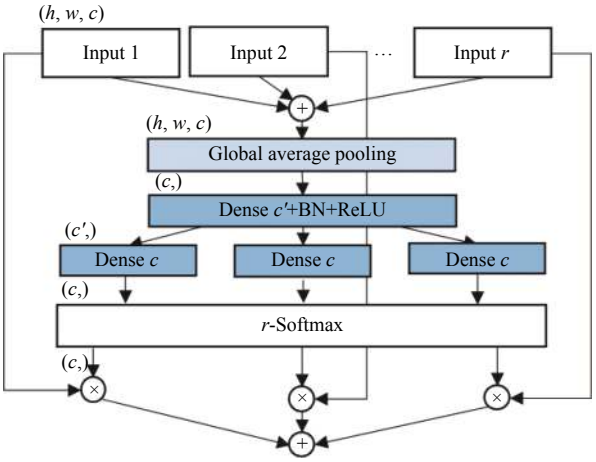

图 3 基组内的片组注意力模块

© 中国科学院软件研究所 http://www.c-s-a.org.cn

设$V^k \in \mathbb{R}^{H\times W\times C/K}$为第$k$个基于通道的软注意力来聚合的基组特征表示, 其中每个分量由片组特征加权组合得到, 如式 (8) 所示, V_c^k为第k个基组的第c个通道分量的表达:

$$V_c^k = \sum_{i=1}^{R} \alpha_i^k(c)\, U_{R(k-1)+i} \tag{8}$$

其中, $\alpha_i^k(c)$表示经过 Softmax 后所得权重, 算法如式 (9):

$$\alpha_i^k(c) = \begin{cases} \dfrac{\exp\left(\mathcal{G}_i^c\left(s^k\right)\right)}{\sum_{j=0}^{R}\exp\left(\mathcal{G}_j^c\left(s^k\right)\right)}, & \text{if } R>1 \\ \dfrac{1}{1+\exp\left(-\mathcal{G}_i^c\left(s^k\right)\right)}, & \text{if } R=1 \end{cases} \tag{9}$$

其中, 权重映射函数$\mathcal{G}$为两个全连接层及一个 ReLU 激活函数 (结构见图 3), $\mathcal{G}_i^c$则通过全局平均池化结果s^k, 为基组内每个片组生成映射权重, 从而生成第c个通道分量的表达.

最后, 我们使用整合函数得到分块中加入了片组注意力映射的整体特征表达:

$$V = \text{Contact}\left(V^1, V^2, \cdots, V^K\right) \tag{10}$$

借鉴 ResNet 的恒等映射机制, 最终分块输出为 Y:

$$Y = V + \mathcal{T}(X) \tag{11}$$

其中, $\mathcal{T}$用于统一残差模块的输出形式, 降低计算成本的同时能增强注意力映射的表达. 在行为识别中, 有效的特征学习是获得高准确率的前提. 通过多路径和恒等映射模块, 片组注意力机制能有效学习特征图层级的注意表达. 实验表明, 片组注意力机制可大幅度提高网络的学习能力, 从而显著地提升了行为识别的准确率.

1.4 时移模块

对于视频分类任务, 传统的 2D CNNs 由于被设计适应基于二维图形的抽象学习, 因此无法做到对视频 (行为) 进行时空建模. 3D CNNs 虽然可以直接对视频进行时空建模, 但其对硬件的计算能力要求较高, 效率较低.

为了能在不增加计算量的前提下提高网络对时空信息的建模能力, 我们在基于时间稀疏分组随机采样策略的双流网络中加入时移模块 (temporal shift module). 以基于瓶颈结构的 ResNet 为例, 我们在每个残差块中插入时移模块, 如图 4 所示.

在基于图像特征的抽取与传递过程中, 网络中的特征图谱通常可以表达为$A \in \mathbb{R}^{N\times C\times T\times H\times W}$, 其中$N$为批处理大小, C为通道数, T 代表时间维度, H 和W则表征空间分辨率. 假设批处理大小为 1, 在时间维度上, 代表不同时刻的向量用不同的颜色表示, 如图 4 所示. 我们在通道维度上对特征进行反向移动, 这同时也表现为在时间维度上进行错位, 这使得相邻帧的信息与当前帧混合在了一起.

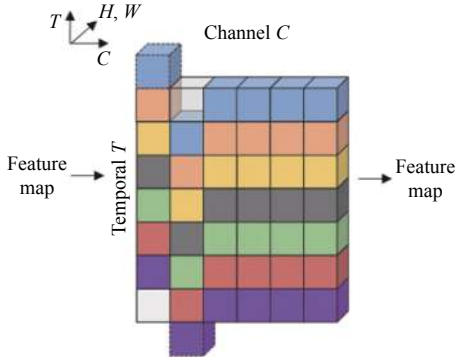

图 4 时移模块

在特征抽取过程中, 时移相当于将卷积分为数据移动和计算两步. 如在 1D 卷积过程中, 设 X 为一维向量, $W=(w_1,w_2,w_3)$为卷积参数, 则卷积过程可表示为:

$$Y_i = Conv(W,X) = w_1X_{i-1} + w_2X_i + w_3X_{i+1} \tag{12}$$

时移操作相当于把式 (12) 分解为两步:

1) 平移置换:

$$X_i^{-1} = X_{i-1}, X_i^0 = X_i, X_i^{+1} = X_{i+1} \tag{13}$$

2) 乘积累加运算:

$$Y = w_1X^{-1} + w_2X^0 + w_3X^{+1} \tag{14}$$

由于行为分析涉及视频帧 (二维图像), 我们把时移模块应用到了 2D 卷积中, 可以看出时移模块相较原始模型不会额外增加计算开销. 在平移置换的具体实践中, 我们将前1/16通道下的特征上移一个步长 (+1), 随后的1/16通道则进行下移 (−1), 剩余通道不移动 (0), 移空的位置用 0 填充. 平移置换相当于在当前帧的特征图谱中用前后帧的信息进行小范围替换, 即每一帧均融合了其前一帧和后一帧的部分特征 (边界除外).

实验表明, 大幅增加平移时的通道比例以增加当前帧中混合的前后帧的特征信息不会提高网络的时序建模能力, 相反会损害网络性能. 这是由于过多的置换会损害当前帧的正常信息表达, 过度的信息交叉对网络学习造成了负面干扰. 同时我们也扩展研究了时移模块的一些变体, 实践发现, 过大的平移幅度如上下移

© 中国科学院软件研究所 http://www.c-s-a.org.cn

动两个步长 (+2、−2) 难以帮助网络优化, 将特征图谱以相同比例在所有时刻上进行置换也无益于性能提升. 基于本文实验, 上下平移1/16的通道来进行时序特征的引入, 能在不增加计算量的前提下提高网络的时空建模能力. 详细实验数据可见第 2 节.

2 实验

2.1 实验设置

实验环境为 PyTorch 1.4.0, 显卡设备为 Tesla V100-SXM2 (显存为 32 GB), 处理器设备为英特尔至强 4110 (2.1 GHz, 8 核), 操作系统为 CentOS 7.5.1804.

为了说明算法的有效性和鲁棒性, 实验数据集包括 UCF101 数据集[16] 和 HMDB51 数据集[24]. 其中 UCF101 共包含 101 类的 13 320 个主要内容为人类体育运动的短视频; HMDB51 则主要来源于网站视频或电影, 共有 51 类人体行为的 6 849 个视频. 对于这两个公共数据集, 本文均使用其官方提供的划分方式 (Split 1) 作为训练计划, 训练集和验证集的比例分别为 2.5:1 (9537:3783) 和 2.3:1 (3570:1530).

在数据预处理阶段, 抽取 RGB 图像和光流图像作为空域和时域特征输入, 同时将数据以多位点随机剪裁的方式 (并调整至 224×224), 结合随机水平翻转 (概率为 0.5) 进行数据增强.

在训练过程中, 采用标准交叉熵损失的学习策略, 在总数为 50 次的迭代中设置前 20 轮的学习率为 0.001, 在第 20 和 40 轮分别降为原来的 0.1 倍, 批处理大小为 90, 动量为 0.9, 分组采样数为 8, 权重衰减为 $5e^{-4}$, Dropout 参数为 0.8, 使用随机梯度下降 (SGD) 对模型参数进行更新.

在测试阶段, 统一在全像素图像上进行左中右方式剪裁以增强测试数据. 并以 1:1.5 的比例拟合 RGB 网络和光流网络的判别分数作为双流 TS-SA 网络的最终结果.

2.2 对比实验

各经典算法准确率对比如表 1 所示, 表 1 中 UCF101 与 HMDB51 数据集下的除本文方法外的数据 (准确率) 均来自于 Wang 等人[17] 的实验. 从表 1 中可以看出, iDT[6] 结合 Fisher Vector 作为最好的传统特征抽取方法之一效果明显, 但在 UCF101 和 HMDB51 上的识别精度可看出其与深度学习方法尚有差距. Two-Stream[15] 作为经典的原始双流网络, 在两个数据集上的识别效果提升明显. C3D[9] 作为更适合学习时空特征的代表网络并没有在精度上超过 Two-Stream, 推测是由于单一地使用 RGB 图像还不能够很好地对外观和运动特征进行统一建模.

表 1 各算法性能对比

算法	准确率 (%)		输入尺寸	视频级计算量 (GFLOPs)	视频级参数量 (M)
	UCF101	HMDB51			
iDT+FV	85.90	57.20	—	—	—
Two-Stream	88.00	59.40	3×224×224	205.5	23.5
C3D	85.20	—	3×16×112×112	38.6	78.1
本文算法	95.00	72.55	3×224×224	35.2	30.4

在视频级计算量 (FLOPs) 上, 均以批处理大小为 1, 视频帧数为 50, 分组采样数为 8 为前提进行 RGB 网络计算量统计. 本文算法在视频级计算量上由于时间稀疏分组随机采样策略优势明显. 因时移模块与分组注意力模块的加入, 模型在参数量上对比原始的网络有小幅增加, 但考虑到模型性能的提升与整体计算量的下降, 本文算法依旧具有较强优势.

2.3 消融实验

为了进一步验证本文提出的策略的优势及有效性, 本文针对算法策略、片段聚合方式以及主干网络的差异进行了消融实验.

为了验证与分析第 1 章中算法策略的有效性及相对重要性, 实验采用 ResNet-50 为主干网络, 在两个数据集上对比了时间稀疏分组随机采样策略 (表 2 中简写为 STGRS)、片组注意力模块 (表 2 中简写为 SA)、时移模块 (表 2 中简写为 TS) 及其组合的准确率, 具体见表 2.

由表 2 可知, 对比密集采样策略, 时间稀疏分组随机采样策略的优势明显, 在不增加计算量的同时成功对行为进行了长时程建模, 在 UCF101、HMDB51 数据集上分别提升了 5.89% 和 3.33% 的识别精度.

在分组策略的基础上, 分别只添加时移模块和注意力模块, 由表 2 可见两种策略在 3 个数据集上均能展现出对网络学习性能的优化. 其中片组注意力模块

© 中国科学院软件研究所 http://www.c-s-a.org.cn

加入的结果令人瞩目, 在 UCF101、HMDB51 数据集上的识别精度分别提升了 6.40%、1.93%, 这说明多路径和特征图谱注意结合的片组注意力机制能在网络中强化学习时的重要特征.

表 2 算法策略识别精度对比 (%)

策略	UCF101			HMDB51		
	RGB网络	光流网络	双流网络	RGB网络	光流网络	双流网络
ResNet-50 (Dense Sampling)	80.91	71.85	89.48	47.58	40.52	59.80
ResNet-50+STGRS	80.73	85.12	92.81	48.76	58.10	65.69
ResNet-50+STGRS+TS	81.63	84.96	92.73	50.07	59.48	67.52
ResNet-50+STGRS+SA	89.40	86.94	94.74	56.73	61.83	72.09
ResNet-50+STGRS+TS+SA	89.61	87.10	95.00	58.56	61.83	72.55

单独添加时移模块较单独增加片组注意力模块的提升较低, 但时移模块与片组注意力模块的组合在 3 个数据集上分别提升了 6.86% 和 2.19%, 说明两种策略的组合能最优化双流网络的识别性能.

接着, 文章试验了片段间不同融合方式对结果的影响. 如表 3 所示, 实验依次比较了最大值融合、平均融合及加权平均融合对精度的影响. 由于平均融合综合考虑了不同时序处的信息, 效果最佳.

表 3 不同融合方式对识别精度的影响 (%)

融合方式	UCF101			HMDB51		
	RGB网络	光流网络	双流网络	RGB网络	光流网络	双流网络
Max	89.45	83.48	94.08	57.45	56.14	71.57
Average	89.61	87.10	95.00	58.56	61.83	72.55
Weighted Average	87.84	83.80	93.29	55.95	49.22	67.25

最后, 由于不同的网络有着不同的学习能力, 一般情况下, 网络越深或越复杂, 意味着其载体容量越大, 所以学习能力越强. 本文对比了不同主干网络下基于两个数据集的识别性能, 结果如表 4 所示.

表 4 不同融合方式对识别精度的影响 (%)

主干网络	UCF101			HMDB51		
	RGB网络	光流网络	双流网络	RGB网络	光流网络	双流网络
SE-ResNet-101	87.55	85.46	94.05	54.97	60.72	70.52
ResNeXt-101	86.55	88.95	94.32	56.73	62.09	71.90
本文算法(TS-SA-ResNet-50)	89.61	87.10	95.00	58.56	61.83	72.55

由表 4 可知, 在保证片段采样数、批处理大小等可控超参数一致的条件下, 本文提出的基于 ResNet-50 的 TA-SA 网络以更轻量的网络结构超过了使用分组卷积改进了的 ResNeXt-101[22] 网络和融合了压缩与激励 (squeeze and excitation, SE) 模块[21] 的 SE-ResNet-101 网络, 充分说明本文算法可以在行为识别任务中实现高效、快速、高准确率的识别效果.

3 结论与展望

本文提出了基于时移和片组注意力融合的时间分组双流深度网络并全面评估了各个模块及其组合的性能. 实验结果表明, 对视频数据进行时间稀疏分组随机采样策略能对行为内容进行长时程高效建模, 且时移模块和片组注意力机制的组合能有效捕获时空特征, 提升网络泛化性能. 相较目前多数行为识别算法, 本文算法在公共数据集中被证明更具有普适性和鲁棒性. 为了进一步提高算法的识别性能, 今后还可从更高效的主干网络优化及多模态特征融合的方向进行深入研究.

参考文献

1 Ben Mabrouk A, Zagrouba E. Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Systems with Applications, 2018, 91: 480–491. [doi: 10.1016/j.eswa.2017.09.029]

2 Dhiman C, Vishwakarma DK. A robust framework for abnormal human action recognition using *R*-transform and Zernike moments in depth videos. IEEE Sensors Journal,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2019, 19(13): 5195 –5203. [doi: 10.1109/JSEN.2019.2903645]

3 Ahmad Z, Khan N. Human action recognition using deep multilevel multimodal (M^2) fusion of depth and inertial sensors. IEEE Sensors Journal, 2020, 20(3): 1445 –1455. [doi: 10.1109/JSEN.2019.2947446]

4 Vyas S, Rawat YS, Shah M. Multi-view action recognition using cross-view video prediction. Proceedings of the European Conference on Computer Vision. Glasgow: Springer, 2020. 427–444.

5 Venkataraman V, Turaga P, Lehrer N, *et al*. Attractor-shape for dynamical analysis of human movement: Applications in stroke rehabilitation and action recognition. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland: IEEE, 2013. 514–520.

6 Wang H, Schmid C. Action recognition with improved trajectories. Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney: IEEE, 2013. 3551–3558.

7 O'Mahony N, Campbell S, Carvalho A, *et al*. Deep learning vs. traditional computer vision. Proceedings of the Science and Information Conference. Las Vegas: Springer, 2019. 128–144.

8 Kay W, Carreira J, Simonyan K, *et al*. The kinetics human action video dataset. arXiv: 1705.06950, 2017.

9 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 4489–4497.

10 Sharma S, Kiros R, Salakhutdinov R. Action recognition using visual attention. arXiv: 1511.04119, 2015.

11 Ng JYH, Hausknecht M, Vijayanarasimhan S, *et al*. Beyond short snippets: Deep networks for video classification. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4694–4702.

12 Li ZY, Gavrilyuk K, Gavves E, *et al*. Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding, 2018, 166: 41–50. [doi: 10.1016/j.cviu.2017.10.011]

13 Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1933–1941.

14 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1725–1732.

15 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014. 568–576.

16 Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv: 1212.0402, 2012.

17 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition, Proceedings of the European Conference on Computer Vision. Floren: Springer, 2016. 20–36.

18 Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 7082–7092.

19 Zhang H, Wu CR, Zhang ZY, *et al*. ResNest: Split-attention networks. arXiv: 2004.08955, 2020.

20 Li X, Wang WH, Hu XL, *et al*. Selective kernel networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 510–519.

21 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-Excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

22 Xie SN, Girshick R, Dollár P, *et al*. Aggregated residual transformations for deep neural networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 5987–5995.

23 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

24 Kuehne H, Jhuang H, Garrote E, *et al*. HMDB: A large video database for human motion recognition. Proceedings of the 2011 International Conference on Computer Vision. Barcelona: IEEE, 2011. 2556–2563.

© 中国科学院软件研究所 http://www.c-s-a.org.cn