基于双流-非局部时空残差卷积神经网络的人体行为识别

钱惠敏 陈 实* 皇甫晓瑛
(河海大学 南京 211100)

摘 要：3维卷积神经网络(3D CNN)与双流卷积神经网络(two-stream CNN)是视频中人体行为识别研究的常用架构，且各有优势。该文旨在研究结合两种架构且复杂度低、识别精度高的人体行为识别模型。具体地，该文提出基于通道剪枝的双流-非局部时空残差卷积神经网络(TPNLST-ResCNN)，该网络采用双流架构，分别在时间流子网络和空间流子网络采用时空残差卷积神经网络(ST-ResCNN)，并采用均值融合算法融合两个子网络的识别结果。进一步地，为了降低网络的复杂度，该文提出了针对时空残差卷积神经网络的通道剪枝方案，在实现模型压缩的同时，可基本保持模型的识别精度；为了使得压缩后网络能更好地学习到输入视频中人体行为变化的长距离时空依赖关系，提高网络的识别精度，该文提出在剪枝后网络的首个残差型时空卷积块前引入一个非局部模块。实验结果表明，该文提出的人体行为识别模型在公共数据集UCF101和HMDB51上的识别准确率分别为98.33%和74.63%。与现有方法相比，该文模型具有参数量小、识别精度高的优点。

关键词：人体行为识别；双流卷积神经网络；3维卷积神经网络；网络剪枝；非局部模块
中图分类号：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2024)03-1100-09
DOI: 10.11999/JEIT230168

Human Activities Recognition Based on Two-stream NonLocal Spatial Temporal Residual Convolution Neural Network

QIAN Huimin CHEN Shi HUANGFU Xiaoying
(*Hohai University*, *Nanjing* 211100, *China*)

Abstract: Three-Dimensional Convolution Neural Network (3D CNN) and two-stream Convolution Neural Network (two-stream CNN) are commonly-used for human activities recognition, and each has its own advantages. A human activities recognition model with low complexity and high recognition accuracy is proposed by combining the two architectures. Specifically, a Two-stream NonLocal Spatial Temporal Residual Convolution Neural Network based onchannel Pruning (TPNLST-ResCNN) is proposed in this paper. And Spatial Temporal Residual Convolution Neural Networks (ST-ResCNN) are used both in the temporal stream subnetwork and the spatial stream subnetwork. The final recognition results are acquired by fusing the recognition results of the two subnetworks under a mean fusion algorithm. Furthermore, in order to reduce the complexity of the network, a channel pruning scheme for ST-ResCNN is presented to achieve model compression. In order to enable the compressed network to learn the long-distance spatiotemporal dependencies of human activity changes better and improve the recognition accuracy of the network, a nonlocal block is introduced before the first residual spatial temporal convolution block of the pruned network. The experimental results show that the recognition accuracies of the proposed human activity recognition model are 98.33% and 74.63% on the public dataset UCF101 and HMDB51, respectively. Compared with the existed algorithms, the proposed model in this paper has fewer parameters and higher recognition accuracy.

Key words: Human activities recognition; Two-stream Convolution Neural Network(two-stream CNN); 3D Convolution Neural Network(3D CNN); Network pruning; NonLocal block

1 引言

视频中的人体行为识别旨在对人体的运动模式进行描述和识别，并最终分析出行为所暗含的情感和目的。其应用前景和场景较为广泛，如基于互联网的视频检索、人机交互、医疗保健、智能安防等等。

目前，基于深度学习的视频中人体行为识别的

收稿日期：2023-03-16；改回日期：2023-07-05；网络出版：2023-07-10
*通信作者： 陈实 1374400532@qq.com

研究受到了广泛关注，并取得了较多的研究成果。最为常用的深度神经网络架构包括3维卷积神经网络(Three-Dimensional Convolution Neural Network, 3D CNN)(简称3D网络)[1-4]、双流卷积神经网络(two-stream Convolution Neural Network, two-stream CNN)(简称双流网络)[5-7]，以及两种架构的融合。3D网络旨在端到端地从视频段中学习人体行为的表征和分类识别，如果直接采用3D网络从视频中学习人体行为表示，由于输入视频段的帧长受限，对空间域轮廓信息和时间域运动信息的学习效果受限，此外，随着视频数据的增加，通常需要更深的3D网络来学习人体行为的特征表示，因而会导致网络的参数量过大而难以训练。鉴于此，研究者提出了许多改进网络，如双流膨胀3D卷积网络(two-stream Inflated 3D convolution network, I3D)、伪3D残差网络(Pseudo-3D residual networks, P3D)等[2,3]。双流网络旨在分别采用2维卷积网络从视频数据的空间流和时间流中学习人体行为的表征和分类识别，并最终融合空间流、时间流的识别结果。但是，空间流和时间流的交互融合、时间流信息的表示和学习方式都有待改进。例如，光流序列常用作视频数据的时间流信息，而采用传统方法计算光流序列不仅需要巨大的计算开销，其生成过程需独立于双流网络之外，从而导致不能实现端到端的人体行为识别。基于此，基于深度神经网络的光流序列提取/生成方法被相继提出，如FlowNet2.0、金字塔扭曲代价容量网络(Pyramid Warping Cost-volumn Network, PWC-Net)等[8,9]。此外，双流网络的子网络通常采用2维卷积运算，不能有效提取视频帧间的时序信息。

3D网络架构和双流网络架构各有优势，本文旨在针对视频中的人体行为识别，研究结合两种架构的、性能更优的双流-时空残差卷积神经网络(Two-stream Spatial Temporal Residual Convolution Neural Network, TST-ResCNN)。这里的性能更优主要指网络的复杂度低、识别精度高。

网络剪枝是压缩网络、降低网络复杂度的常用方法之一。其中，通道剪枝是网络结构化剪枝的方法之一，旨在删除不重要的通道，从而加快网络的推理速度。本文将针对时空残差卷积神经网络设计基于通道剪枝的网络压缩方案。但是，网络压缩通常会降低网络的精度。本文还将进一步研究如何有效保持网络的识别精度。

视频中的人体行为在视频帧序列的时、空域上均有信息变化，因此，网络若能对连续的视频帧之间的信息变化进行有效学习，如人体姿态的变化、人与人之间/人与物体之间的位置变化等长距离时空依赖关系，其识别精度将被提高。例如，Wei等人[10]发现通过堆叠深度卷积网络中的卷积运算，扩大感受野，可以捕捉输入的长距离依赖关系。然而，由此导致的网络深度增加会增加网络的训练难度，并且其中的重复局部运算会导致部分局部重要信息的丢失。Li等人[11]提出时间差分网络(Temporal Difference Networks, TDN)，通过分别设计短时差分模块和长时差分模块，分别实现对短时和长时运动中变化信息的时序建模，由此提高网络对长距离依赖信息的学习能力。但这种模块难以直接运用在现有人体行为识别模型。受到图像去噪领域的非局部均值启发，研究者发现，非局部模块可以对长距离依赖关系建模[12]。因此，本文通过在网络中引入非局部模块，计算多帧输入特征图中所有位置像素间的相关性，表达信息变化的长距离时空依赖，并为相关性更大的位置分配更大的注意力权重，提高后续网络对变化信息的学习能力，从而提高网络的识别精度。并且，在原网络中引入非局部模块，不需要改变网络的整体结构以及网络的输入方式，结构清晰，参数量小。

综上，本文通过设计针对时空残差卷积神经网络的通道剪枝方案，并在网络中引入非局部模块，提出基于通道剪枝的双流-非局部时空残差卷积神经网络(Two-stream NonLocal Spatial Temporal Residual Convolution Neural Network based on channel Pruning , TPNLST-ResCNN)，本网络具有复杂度低、识别准确率高的优点。本文的主要贡献包括：

(1) 提出基于通道剪枝的双流-非局部时空残差卷积神经网络(TPNLST-ResCNN)，本网络结合了3D网络和双流网络的优势，具有参数量少而识别精度较高的特性；

(2) 提出针对残差型网络结构的通道剪枝方案，实现了深度神经网络的模型压缩，降低了模型的训练难度；

(3) 提出在时间流子网络和空间流子网络的首个残差型时空卷积块前增加一个非局部模块，提高网络对人体行为的长距离时空依赖关系的学习能力，提高网络的识别精度。

2　基于通道剪枝的双流-非局部时空残差卷积神经网络

鉴于双流网络和3D网络在识别视频中的人体行为时的各自优势，本文提出融合双流架构和3D架构的双流-非局部时空残差卷积神经网络，网络模型具体如图1所示。

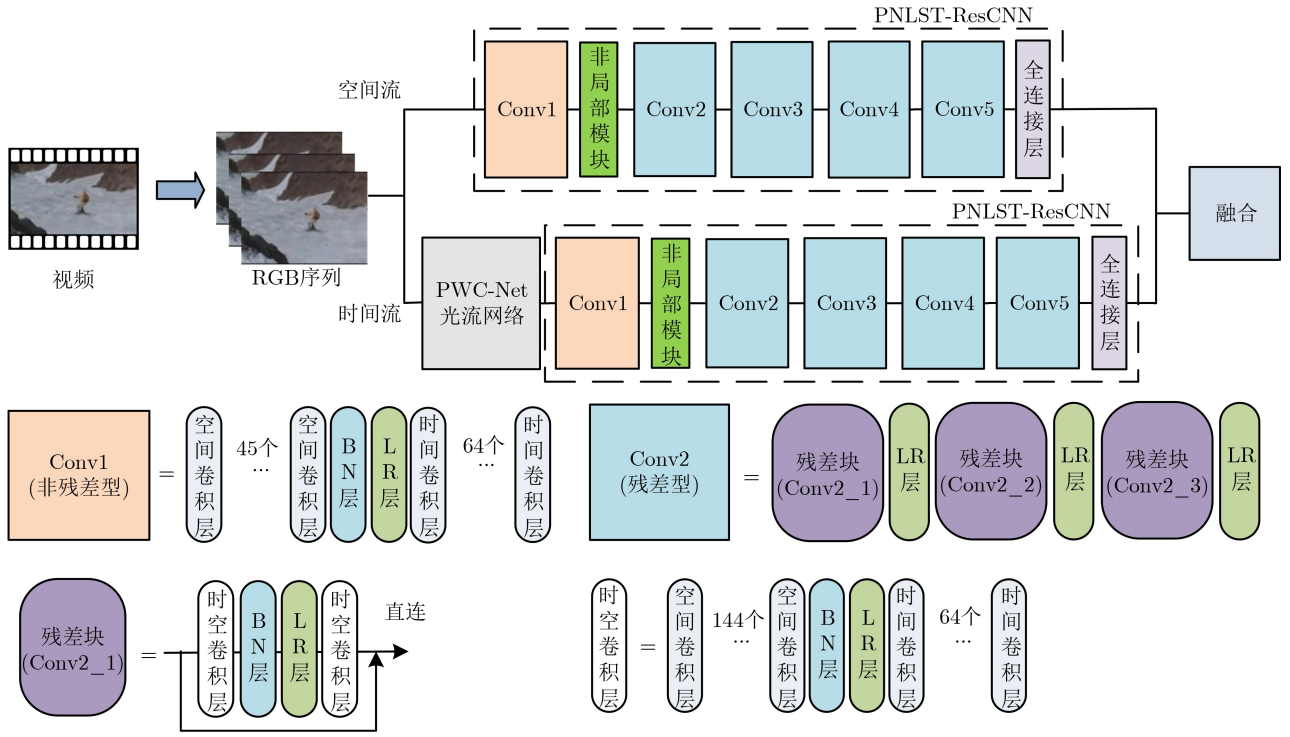


图 1 双流-非局部时空残差卷积神经网络

本网络分别从视频的时间流和空间流中学习人体行为的表征和分类，并融合双流结果实现人体行为的识别。首先，给定待识别视频段，以按帧提取方式生成三原色(Red Green Blue, RGB)序列；然后，将RGB序列分别送入空间流子网络和时间流子网络，两个子网络中均采用基于通道剪枝后的非局部时空残差卷积神经网络(NonLocal Spatial-Temporal Residual Convolution Neural Network based on channel Pruning, PNLST-ResCNN)实现人体行为的特征表示与分类；最后，融合空间、时间流子网络的结果，得到双流-非局部时空残差卷积神经网络的识别结果。

本文采用的时空残差卷积神经网络(Spatial Temporal Residual Convolution Neural Networks, ST-ResCNN)参考了文献[13]提出的网络结构，该网络将3D卷积核进行时空分解，同时通过保证空间卷积核的个数保持网络的参数量，确保不损失网络的信息表达能力。为了降低网络复杂度，提高识别精度，本文以ST-ResCNN为基础，提出PNLST-ResCNN，其网络结构如图1所示。PNLST-ResCNN包含5个时空卷积块(Conv1～Conv5)、1个非局部模块和1个全连接层，其中，Conv1为非残差型时空卷积块，Conv2～Conv5为残差型时空卷积块。这4个残差型时空卷积块均包含多个残差块，为了方便叙述，残差型时空卷积块分别记为Conv2_x(x=1,2,3)，Conv3_x(x=1,2,3,4)，Conv4_x(x=1,2,3,4,5,6)，Conv5_x(x=1,2,3)，x表示残差块编号。每个残差块又由时空卷积层、批规范层(BatchNorm, BN)、激活函数层(Leaky ReLU, LR)、时空卷积层及直连结构组成。图1给出了Conv2_1的结构。更进一步地，时空卷积层由空间卷积层，BN,LR激活函数层和时间卷积层组成。文献[13]给出了具体的非/残差型时空卷积块的结构。特别地，时间流子网络先采用金字塔扭曲代价容量(Pyramid Warping Cost-volumn Network, PWC-Net)光流提取网络从RGB序列中提取具有运动信息的光流序列[9]；然后，采用PNLST-ResCNN实现基于光流序列的人体行为特征表示与分类。

本文采用的双流-时空残差卷积神经网络(TST-ResCNN)的参数量为：光流图像估计网络PWC-Net的参数量约为8.75M，单流识别网络ST-ResCNN的参数量约为63.5M，整体网络的参数量约136M。因此，网络的训练和使用均对硬件和数据集的大小有较高要求。鉴于ST-ResCNN主要由残差型时空卷积块构成，本文将针对其中的残差型时空卷积块设计通道剪枝方案，在尽可能保持精度的条件下压缩网络，减少网络参数量。

2.1 通道剪枝

通道剪枝是指对网络中不重要的卷积通道进行裁剪以降低网络复杂度。文献[14]指出，卷积网络中BN层的缩放因子可度量其前继卷积通道的重要性，裁剪不重要的卷积通道即可实现网络压缩。具

体地，在网络训练过程中，通过在损失函数中引入惩罚项对网络进行稀疏化训练；继而根据网络BN层中与每一个卷积通道对应的缩放因子确定卷积通道的重要性，并对缩放因子值小于给定剪枝阈值的通道进行裁剪；最后，对剪枝后的网络进行恢复性训练。

假设原网络的损失函数为交叉熵损失L

$$L = -\sum_{k} Y_k \ln\left(f\left(x_k, \boldsymbol{W}\right)\right) \tag{1}$$

其中，Y_k表示输入x_k的真实标签；$f(x_k, \boldsymbol{W})$表示输入x_k经网络(参数矩阵为$\boldsymbol{W}$时)的预测标签。网络稀疏化训练时，引入由BN层缩放因子γ确定的L1正则化惩罚项，此时，损失函数L_p为

$$L_p = L + \zeta \sum_{\gamma \in \Gamma} |\gamma| \tag{2}$$

其中，ζ为稀疏因子，Γ是网络中所有BN层缩放因子的集合。

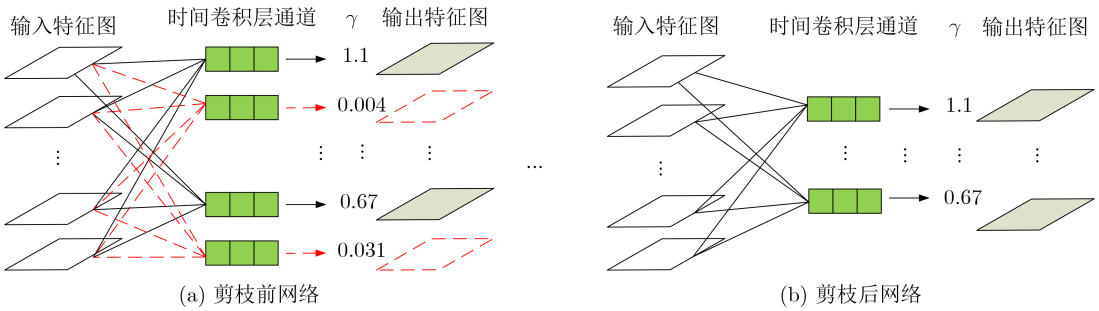
图2为ST-ResCNN经稀疏化训练后的时间卷积层的通道剪枝示意图。如图2所示，时间卷积层的每个通道均有与其一一对应的BN层缩放因子γ，根据设定的剪枝阈值，将γ小于阈值的时间卷积通道进行裁剪。假设某个BN层缩放因子$\gamma = 0.004$，小于设定的剪枝阈值，则将连接该BN层的前继卷积通道的权重、其输入和输出连接(即图中虚线部分)一并删除，由此得到结构更为紧凑的剪枝后网络

由此可知，损失函数L_p中稀疏因子ζ的大小决定了缩放因子γ的稀疏程度，而γ的稀疏程度不同，网络的剪枝上限也不尽相同，这进而会影响剪枝后网络经恢复性训练后的性能。本文认为一个合适的稀疏因子应满足两个要求：(1)为了提高网络的压缩率，γ的稀疏程度不能太低，即网络经稀疏化训练后γ的值在0附近处的数量应占比50%左右；(2)为了剪枝后网络经恢复性训练后的性能较好，稀疏程度不能太高，即网络剪枝后剩余的γ值在0附近处的数量不能太多。本文通过实验确定时间流子网络和空间流子网络的稀疏因子。以HMDB51数据集的空间流子网络为例，统计ζ分别取0, 10^{-5}, 10^{-4}, 10^{-3}时，稀疏化训练后γ的分布，通过分析选择稀疏因子为10^{-4}。同理，时间流子网络的稀疏因子取为10^{-5}。

2.2 残差型时空卷积块的通道剪枝方案

本文将采用基于BN层缩放因子的通道剪枝方法对子网络ST-ResCNN进行压缩。为了保持ST-ResCNN网络的性能优势，提出两个剪枝原则：(1)剪枝时不能破坏其残差结构；(2)剪枝时不能破坏残差块中空间卷积层的结构。遵循剪枝原则1，保留每一个残差块的直连结构，参照图1中残差块的结构，不能对残差块的最后一个时间卷积层进行剪枝。遵循剪枝原则2，仅对每个残差型时空卷积块中的时间卷积层进行剪枝。为了进一步压缩网络，本文提出通过减少输入全连接层的通道数的方法实现全连接层参与剪枝。也就是说，删除最后一个残差块(Conv5_3)的直连，并对Conv5_3的最后一个时间卷积层也进行剪枝。

综上，本文对ST-ResCNN的剪枝方案为：删除网络的最后一个残差块(即Conv5_3)的直连，并将该残差块中所有的时间卷积层参与剪枝，如图3(b)所示，红框标注的时间卷积层均参与剪枝；其余的残差型时空卷积块中的所有残差块，包括Conv2_x (x=1,2,3), Conv3_x (x=1,2,3,4), Conv4_x (x=1,2,3,4,5,6), Conv5_x (x=1,2)，均保留其直连，对除残差块的最后一个时间卷积层之外的所有时间卷积层都进行剪枝。以Conv5_2为例，仅红框标注的第1个时间卷积层参与剪枝，如图3(a)所示。需要说明的是，剪枝方案将BN层的缩放因子作为衡量通道重要性的指标，并根据该指标剪除对网络贡献小的非重要通道，从而压缩网络。同时，正因为被剪除通道对网络的贡献小，剪枝后网络的识别精度损失不大。例如，在UCF101数据集上，模型经过剪枝后，压缩率约为45%，而识别精度仅降低了0.05%。



(a) 剪枝前网络　　(b) 剪枝后网络

图 2 时间卷积层的通道剪枝示意图

3 基于通道剪枝的非局部时空残差卷积神经网络

Varol等人[15]发现提高双流网络的输入视频段的帧长，有助于提高网络精度。本文实验也验证了这一点。例如，本文在HMDB51数据集上，采用双流-时空残差卷积神经网络进行识别，当将输入帧长从8提高到16时，网络的识别精度提高了7.1%，但是，本文在对经通道剪枝后的双流-时空残差卷积神经网络执行相同实验时发现，提高输入帧长并不能明显提高网络的识别精度。因此，本文认为通道剪枝降低了原网络对人体行为在长时间段内的变化信息的学习能力。文献[12]提出非局部模块可以捕获图像、视频中的长距离依赖信息。受此启发，本文将非局部模块引入通道剪枝后的双流-时空残差卷积神经网络中，以提高网络对视频中长距离依赖信息的学习能力。

3.1 非局部模块

假设非局部模块的输入特征图为I_1,I_2，x_i和x_j分别为I_1,I_2在位置i和j处的特征值，Z为非局部模块的输出，且Z_i与输入特征图I_1中的x_i对应，则

$$Z_i = \boldsymbol{W}_Z \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) + x_i \tag{3}$$

其中，$f(x_i, x_j)$为输入特征图I_1,I_2中位置i和j处的特征值的相关性度量，本文中函数f选用式(4)所示的嵌入式高斯函数[12]

$$f(x_i, x_j) = \mathrm{e}^{\boldsymbol{\Theta}^{\mathrm{T}}(x_i)\boldsymbol{\Phi}(x_j)} \tag{4}$$

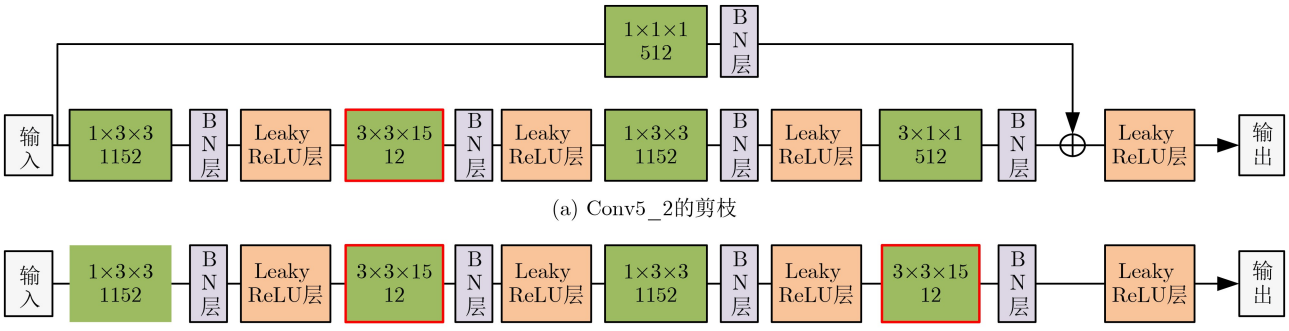
其中，$\boldsymbol{\Theta}(x_i) = \boldsymbol{W}_{\boldsymbol{\Theta}} x_i$，$\boldsymbol{\Phi}(x_j) = \boldsymbol{W}_{\boldsymbol{\Phi}} x_j$，且$\boldsymbol{W}_{\boldsymbol{\Theta}}\boldsymbol{W}_{\boldsymbol{\Phi}}$为映射矩阵，T指转置操作。嵌入式高斯函数先将特征值x_i和x_j分别映射到新的线性空间，再计算它们之间的相关性度量。f的值越小，表示位置i和位置j之间的像素相关性越小。

此外，$\boldsymbol{g}(x_j) = \boldsymbol{W}_{\boldsymbol{g}} x_j$，$\boldsymbol{W}_{\boldsymbol{g}}$和$\boldsymbol{W}_Z$为权重矩阵。$C(x)$为归一化参数，且

$$C(x) = \sum_{\forall j} f(x_i, x_j) \tag{5}$$

需要说明的是，式(3)中的矩阵$\boldsymbol{W}_{\boldsymbol{\Theta}}$, $\boldsymbol{W}_{\boldsymbol{\Phi}}$, $\boldsymbol{W}_{\boldsymbol{g}}$和$\boldsymbol{W}_Z$均由网络训练确定。

图4给出了非局部模块的具体结构。本文中，非局部模块的输入是多帧特征图，输入特征图序列输入模块$\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{g}$(分别对应$\boldsymbol{\Theta}(\cdot)$, $\boldsymbol{\Phi}(\cdot)$和$\boldsymbol{g}(\cdot)$)进行处理，模块f为式(4)所示的嵌入式高斯函数，计算得到每帧输入像素与其他所有帧中像素的相关性权重后，再由1×1×1模块进行维度转换，得到多帧特征图之间的长距离依赖度量矩阵，最后与直连的多帧输入特征图相加，得到非局部模块的输出。



(a) Conv5_2的剪枝

(b) Conv5_3的剪枝

图 3 剪枝方案示意图

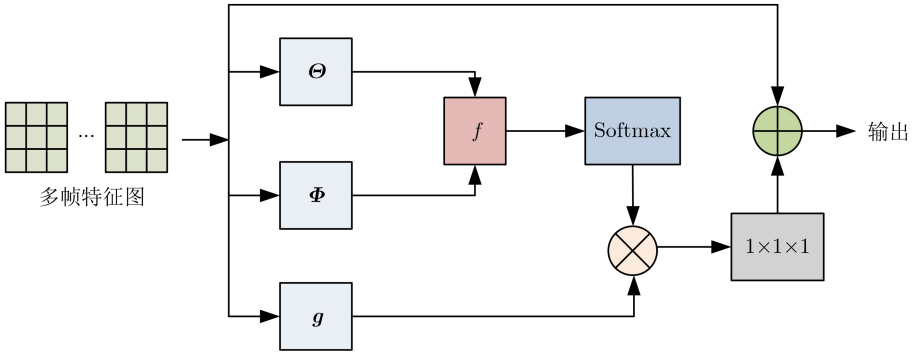


图 4 非局部模块的网络结构

3.2 基于通道剪枝的非局部时空残差卷积神经网络

本文将在时空残差卷积神经网络中引入非局部模块。经分析，为了使得后续网络层能提取到更丰富的语义信息，在网络的靠前层引入非局部模块效果更佳，后续实验也验证这一分析结果。其次，非局部模块的计算需要较大的显存开销，在引入时，数量应该适当。因此，本文考虑在ST-ResCNN的Conv1后增加一个非局部模块，如图1所示。非局部模块的引入，使得改进后的模型具备了对长距离依赖关系的学习能力，此时再进行基于BN层缩放因子的通道剪枝，可确保剪枝后网络能够在提高视频输入的帧长时，进一步提高网络的识别精度。

4 实验及结果分析

4.1 数据集和实验设置

目前常用的人体行为识别视频数据集有UCF101, HMDB51, Kinetics-400, Kinetics-700等[16-20]。本文选择其中较为常用的，样本数量相对较少的两个数据集：UCF101和HMDB51，以验证提出模型在无大数据支撑下的识别效果。

基于数据集，本文采用按帧提取的方式从视频样本中获取RGB图像序列，设置帧提取率为30，并基于PWC-Net生成光流图像序列。训练前，对已有数据集中的样本按训练集:测试集等于7:3划分，为了提高模型的泛化性，对数据集做一定的数据增强。首先，将原始大小为320×240的图像缩放至171×128后随机裁剪为112×112大小；然后将图像以0.5的概率水平翻转。为了降低网络的训练难度，提高网络性能，本文使用了在Kinetics-400上的预训练模型[1]。模型训练时，输入网络的是一系列设定帧长的连续RGB帧或光流序列，并采用基于梯度中心化算法改进的带动量的随机梯度下降算法(Gradient Centralization Stochastic Gradient Descent with Momentum, GC-SGDM)优化器，设置权重衰减为0.000 5，动量为0.9，初始学习率为0.000 1，并以损失是否下降为指标更新学习率，学习耐心设置为10，且根据现有实验条件，设置网络的输入帧长为8时批尺寸为10，输入帧长为16时批尺寸为5，对于UCF101数据集，设置训练epoch为600，HMDB51数据集设置为800。实验硬件为：两块型号为RTX 2080Ti的GPU，型号为i7-7800X @3.50GHz×12的CPU，实验环境为Ubuntu 16.04, CUDA 8.0, CUDNN 7.4，所有实验均在PyTorch框架下完成。

4.2 基于通道剪枝的双流-时空残差卷积神经网络的性能分析

4.2.1 主干网络的深度选择

一般而言，神经网络的深度越深，特征表达能力越强，模型性能越好。但是层数越深的神经网络对训练数据的数量要求也会增加。层数深的神经网络在小数据集上训练时易出现过拟合现象。因此，需要探讨不同大小数据集适用的主干网络的深度。以空间流子网络为例，实验首先采用34层ST-ResCNN作为主干网络。实验发现，在UCF101上，34层ST-ResCNN的性能较优，而在HMDB51上，34层ST-ResCNN出现了过拟合现象。

因此，本文将首先为HMDB51数据集选择合适的网络深度。具体地，本文对4个残差型时空卷积块进行删减，经过试验共得A, B, C, D 4种模型，表1给出了它们的网络层数，每个时空卷积块中的残差块个数，删减后模型的参数量，以及将其作为空间流子网络时的识别精度。在后续针对HMDB51数据集的实验中，本文将使用识别精度最高的模型A。

4.2.2 基于双流-时空残差卷积神经网络的人体行为识别结果

由上可知，在数据集UCF101和HMDB51上，本文将分别采用34层ST-ResCNN、10层ST-ResCNN(即表1中的模型A)作为时间流、空间流子网络的主干网络。设置两个子网络的输入帧长均为16，且34层ST-ResCNN使用在Kinetics-400上的预训练模型。分别采用均值融合法和最大值融合法，融合时间流、空间流子网络的识别结果。实验结果见表2。由表2可知，本文提出的双流-时空残差卷积神经网络在采用均值融合方法时，在UCF101和HMDB51上均获得了较高的识别精度，分别为98.00%和69.20%。

表 1　不同网络深度ST-ResCNN的结构及其识别精度

网络模型	网络层数	参数量(M)	Conv2(个)	Conv3(个)	Conv4(个)	Conv5(个)	精度(%)
A	10	14.38	1	1	1	1	57.70
B	12	15.26	1	2	1	1	55.65
C	12	17.92	1	1	2	1	55.80
D	12	28.54	1	1	1	2	55.17

1 https://github.com/open-mmlab/mmaction

表 2 融合实验结果对比(%)

数据集	空间流	时间流	最大值融合	均值融合
UCF101	94.60	85.67	97.70	98.00
HMDB51	58.63	50.15	62.80	69.20

4.2.3 基于通道剪枝的双流-时空残差卷积神经网络的人体行为识别结果

由2.1.1节可知，根据实验，识别UCF101数据集的时间流和空间流子网络的稀疏因子选为10^{-4}；识别HMDB51数据集的时间流、空间流子网络的稀疏因子分别选为10^{-5}和10^{-4}。此外，网络的压缩率与剪枝阈值相关，剪枝阈值设置较大，压缩后模型更小，但其识别精度也会降低更多。因此，以UCF101数据集为例，本文分别选取剪枝阈值为70%和80%时，对比两个数据集下，剪枝后网络的参数量、模型大小、压缩率和识别精度。根据在模型压缩率和模型精度间折衷的原则，本文选取出不同数据集下子网络的最佳剪枝阈值，如表3所示。实验中，设置网络的输入帧长为8，且使用Kinetics-400数据集上的预训练模型。

由表3可知，在数据集UCF101上，当空间流子网络的剪枝阈值为70%时，剪枝后网络可获得最高精度92.13%，对比表2与原始网络相比仅低了0.95%，而此时模型的压缩率有41.70%。而在HMDB51上，剪枝后的网络获得比剪枝前网络更高的识别精度。分析可知，模型A在HMDB51上仍存在过拟合，经剪枝压缩网络后，性能得到提升。

同时，表3给出了在两个数据集上，采用均值融合算法的识别结果。此外，本文对全连接层是否参与剪枝的模型进行了实验对比，由实验可知，在相同剪枝阈值下，全连接层参与剪枝后网络的模型压缩率提高，且识别精度也有所提高。以34层ST-ResCNN的空间流子网络为例，剪枝阈值为80%时，全连接层参与剪枝后，模型的参数量减少3.86M，模型大小降低15.5MB，压缩率提高6%，且识别精度提高了0.84%。

4.3 基于通道剪枝的双流-非局部时空残差卷积神经网络的性能分析

本文将网络的输入帧长从8提高到16，实验发现，提高输入帧长，在剪枝前网络上能获得更高的精度收益，而剪枝后网络的精度收益较小，如表4所示。本文认为通道剪枝降低了原网络对人体行为在长时间段内的变化信息的学习能力。因此，本文将通过在网络中增加非局部模块来提升网络对视频中长距离信息的学习能力。

在网络中引入非局部模块时，对网络的参数量影响不大。例如，表1中的模型A，其参数量为14.38M，按40%的剪枝阈值剪枝后的模型参数量为8.92M，引入一个非局部模块后网络的参数量为10.05M。但是，非局部模块的计算需要较大的显存开销，对硬件设备的要求较高，因此，本文尝试在网络中引入1个非局部模块，进行相关实验探讨其引入位置。以时间流子网络在HMDB51数据集上的实验为例，分别在10层ST-ResCNN的4个残差型时空卷积块的输入前端添加一个非局部模块进行实验。发现非局部层的位置越靠网络的输入端，测试的精度逐渐变高。最高在第1个残差型时空卷积块前添加非局部模块，比原始网络高出0.68%，最低在第4个残差型时空卷积块前添加非局部模块，只比原始网络高出0.04%。基于此，在不同深度的ST-ResCNN中的第1个残差型时空卷积块前添加非局部模块，并采用上述剪枝方案将网络参与剪枝。再次训练分析其性能。选择输入帧长为16，各子网络的剪枝阈值均为上述表3所示的最佳方案，并将网络进行融合实验。如表5所示，在剪枝后网络中引入非局部模块后，提高网络输入帧长，网络的识别精度增长有提高。

表 3 UCF101和HMDB51上剪枝的实验结果(%)

数据集	子网络	剪枝阈值	模型压缩	精度	融合精度
UCF101	空间流	70	41.70	92.13	96.83
	时间流	80	41.70	81.96	
HMDB51	空间流	40	37.97	59.11	72.27
	时间流	30	27.89	54.97	

表 4 提高输入帧长后网络的识别精度对比

数据集	输入帧长	剪枝前的精度(%)	剪枝后的精度(%)
UCF101	8	96.88	96.83
	16	98.00	97.75
HMDB51	8	62.10	72.27
	16	69.20	73.01

4.4 本文算法与其他算法的比较

本文比较了本文算法与现有算法在相同数据集

表 5 3种网络的对比实验(输入帧长为16、均值融合)

网络名称	数据集	参数量(M)	精度(%)	精度变化(%)
ST-ResCNN	HMDB51	28.76	69.20	+7.10
	UCF101	127.08	98.00	+1.12
PST-ResCNN	HMDB51	19.29	73.01	+0.74
	UCF101	70.29	97.75	+0.92
PNLST-ResCNN	HMDB51	20.11	74.63	+1.53
	UCF101	71.68	98.33	+4.67

表 6　本文算法与现有算法的比较

算法	输入	预训练数据集	参数量(M)	精度(%)	
				UCF101	HMDB51
C3D[2]	RGB	Sports-1M	61.63	82.3	56.8
P3D[5]	RGB	Sports-1M	–	88.6	–
R3D-34[21]	RGB	Kinetics-700	63.52	88.8	59.5
R(2+1)D-50[21]	RGB	Kinetics-700+Sports1M	53.95	93.4	69.4
CIDC[11]	RGB	–	103.00	97.9	75.2
ActionCLIP[22]	RGB	网络数据	85.58	97.1	76.2
STM(ResNet-50)[23]	RGB	ImageNet+Kinetics	–	96.2	72.2
TDN(ResNet-50) [11]	RGB	ImageNet+Kinetics	–	97.4	76.3
R(2+1)D-34[24]	双流	Sports-1M	127.08	95.0	72.7
本文PNLST-ResCNN-34	双流	Kineticts-400	71.68	98.3	–
本文PNLST-ResCNN-10	双流	–	20.11	–	74.6

上的识别结果，如表6所示。需要说明的是，与本文相比较的算法的技术重点均为网络结构改进。除ActionCLIP采用了Transformer架构外[22](预训练数据来源于网络无具体名称)，表6中“输入”仅为“RGB”的行为识别算法均采用3D网络，“输入”为“双流”的行为识别算法均采用RGB+光流作为输入的双流网络。

由表6可知，相对于单流的R(2+1)D-50而言[21]，与双流网络相结合的R(2+1)D-34算法更优[24]，本文算法剪枝前后的结果与针对网络结构进行优化改进的最新技术(ActionCLIP)及采用STM的方法相比[22,23]，在UCF101数据集上，识别率均更高。与TDN相比[11]，本文算法在UCF101数据集上的准确率高了0.93%；虽然TDN在HMDB51数据集上的准确率比本文提出方法高出1.67%，但其主干网络为ResNet-50(参数量为25.5M)，而本文PNLST-ResCNN-10的主干网络只有10层(参数量为10.5M)，因此本文算法具有一定优势。综上，本文提出模型与前沿技术均有可比较性且在中小型数据集上都具有良好的精度，在UCF101和HMDB51数据集上的最高识别精度为98.33%和74.63%。

5　结束语

本文首先针对3D结构和双流结构的各自优缺点，提出双流-时空残差卷积神经网络。为了降低模型复杂度，减轻网络的训练难度，提出针对残差型网络结构的通道剪枝方案，实现网络的压缩。进一步地，本文提出在时间流和空间流网络的首个残差型时空卷积块前增加非局部模块，经剪枝后得到基于通道剪枝的双流-非局部时空残差卷积神经网络(TPNLST-ResCNN)，适当提高了剪枝后网络的复杂度，有利于网络提取长时间段内人体动作的变化特征。除此之外，在未来工作中还会使用更多的图像增强技术来扩大小数据集的规模，也会尝试特征融合代替分数融合以提高实验精度。

参 考 文 献

[1] 白静, 杨瞻源, 彭斌, 等. 三维卷积神经网络及其在视频理解领域中的应用研究[J]. 电子与信息学报, 2023, 45(6): 2273–2283. doi: 10.11999/JEIT220596.
BAI Jing, YANG Zhanyuan, PENG Bin, *et al*. Research on 3D convolutional neural network and its application to video understanding[J]. *Journal of Electronics & Information Technology*, 2023, 45(6): 2273–2283. doi: 10.11999/JEIT220596.

[2] CARREIRA J and ZISSERMAN A. QUO Vadis, action recognition? A new model and the kinetics dataset[C]. The 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6299–6308. doi: 10.1109/CVPR.2017.502.

[3] QIU Zhaofan, YAO Ting, and MEI Tao. Learning spatio-temporal representation with pseudo-3D residual networks[C]. The 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 5534–5542. doi: 10.1109/ICCV.2017.590.

[4] 王粉花, 张强, 黄超, 等. 融合双流三维卷积和注意力机制的动态手势识别[J]. 电子与信息学报, 2021, 43(5): 1389–1396. doi: 10.11999/JEIT200065.
WANG Fenhua, ZHANG Qiang, HUANG Chao, *et al*. Dynamic gesture recognition combining two-stream 3D convolution with attention mechanisms[J]. *Journal of Electronics & Information Technology*, 2021, 43(5): 1389–1396. doi: 10.11999/JEIT200065.

[5] PANG Chen, LU Xuequan, and LYU Lei. Skeleton-based

action recognition through contrasting two-stream spatial-temporal networks[J]. *IEEE Transactions on Multimedia*, 2023, 1520–9210.

[6] VARSHNEY N and BAKARIYA B. Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple CNN streams[J]. *Multimedia Tools and Applications*, 2022, 81(29): 42117–42129. doi: 10.1007/s11042-021-11220-4.

[7] LI Bing, CUI Wei, WANG Wei, *et al*. Two-stream convolution augmented transformer for human activity recognition[C]. The 35th AAAI Conference on Artificial Intelligence, 2021: 286–293.

[8] ILG E, MAYER N, SAIKIA T, *et al*. Flownet 2.0: Evolution of optical flow estimation with deep networks[C]. The 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1647–1655. doi: 10.1109/CVPR.2017.179.

[9] SUN Deqing, YANG Xiaodong, LIU Mingyu, *et al*. PWC-net: CNNs for optical flow using pyramid, warping, and cost volume[C]. The 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8934–8943. doi: 10.1109/CVPR.2018.00931.

[10] WEI S E, RAMAKRISHNA V, KANADE T, *et al*. Convolutional pose machines[C]. The 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 4724–4732. doi: 10.1109/CVPR.2016.511.

[11] LI Xinyu, SHUAI Bing, and TIGHE J. Directional temporal modeling for action recognition[C]. The 16th European Conference on Computer Vision, Glasgow, UK, 2020: 275–291.

[12] WANG Xiaolong, GIRSHICK R, GUPTA A, *et al*. Non-local neural networks[C]. The 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7794–7803. doi: 10.1109/CVPR.2018.00813.

[13] HUANG Min, QIAN Huimin, HAN Yi, *et al*. R(2+1)D-based two-stream CNN for human activities recognition in videos[C]. The 40th Chinese Control Conference, Shanghai, China, 2021: 7932–7937.

[14] LIU Zhuang, LI Jianguo, SHEN Zhiqiang, *et al*. Learning efficient convolutional networks through network slimming[C]. The 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2755–2763. doi: 10.1109/ICCV.2017.298.

[15] 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.

[16] SOOMRO K, ZAMIR A R, and SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[OL]. arXiv preprint arXiv: 1907.06987, 2012.

[17] KUEHNE H, JHUANG H, GARROTE E, *et al*. HMDB: A large video database for human motion recognition[C]. The 2011 International Conference on Computer Vision, Barcelona, Spain, 2011: 2556-2563. doi: 10.1109/ICCV.2011.6126543.

[18] KAY W, CARREIRA J, SIMONYAN K, *et al*. The kinetics human action video dataset[OL]. arXiv preprint arXiv: 1705.06950, 2017.

[19] CARREIRA J, NOLAND E, HILLIER C, *et al*. A short note on the kinetics-700 human action dataset[OL]. arXiv preprint arXiv: 1907.06987, 2019.

[20] KARPATHY A, TODERICI G, SHETTY S, *et al*. Large-scale video classification with convolutional neural networks[C]. The 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1725–1732. doi: 10.1109/CVPR.2014.223.

[21] KATAOKA H, WAKAMIYA T, HARA K, *et al*. Would mega-scale datasets further enhance spatiotemporal 3D CNNs?[OL]. arXiv preprint arXiv: 2004.04968, 2020.

[22] WANG Mengmeng, XING Jiazheng, and LIU Yong. ActionCLIP: A new paradigm for video action recognition[J]. arXiv preprint arXiv: 2109.08472, 2021.

[23] JIANG Boyuan, WANG Mengmeng, GAN Weihao, *et al*. STM: SpatioTemporal and motion encoding for action recognition[C]. The 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 2000–2009. doi: 10.1109/ICCV.2019.00209.

[24] TRAN D, WANG Heng, TORRESANI L, *et al*. A closer look at spatiotemporal convolutions for action recognition[C]. The 2018 IEEE/CVF conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6450–6459. doi: 10.1109/CVPR.2018.00675.

钱惠敏：女，副教授，硕士生导师，研究方向为智能视频监控系统、视频中的人体行为分析、深度学习等.

陈 实：男，硕士生，研究方向为视频中的人体行为分析.

皇甫晓瑛：女，硕士生，研究方向为视频中的人体行为分析.

责任编辑：余 蓉