文章编号：2095-2163(2021)07-0043-07 中图分类号：TP391.41 文献标志码：A

基于协同时空建模的视频行为识别算法

郑 阳，张旭东
（合肥工业大学 计算机与信息学院，合肥 230601）

摘 要：行为识别是计算机视觉领域的一个重要研究方向，已被广泛应用于视频监控、人群分析、人机交互、虚拟现实等领域。而时空建模是视频行为识别的一个重要部分，有效地进行时空建模可以极大地提高行为识别的精度。现有的先进算法采用 3D CNN 学习强大的时空表示，但在计算上是复杂的，这也使得相关部署昂贵；此外，改进的具有时间迁移操作的 2D CNN 算法也被用来进行时空建模，这种算法通过沿时间维度移动一部分特征通道用以进行高效的时序建模。然而，时间迁移操作不允许自适应地重新加权时空特征。以前的工作没有考虑将这两种方法结合利用起来，取长补短，以便更好地建模时空特征。本文提出了一个协作网络用以有效地结合 3D CNN 和 2D 卷积形式的时间迁移模块。特别是一个新的嵌入注意力机制的协同时空模块（Collaborative Spatial-temporal module，CSTM）被提出用以有效的学习时空特征。本文在与时序相关的数据集（Something-Something v1，v2，Jester）上验证了该算法的有效性，并且获得了竞争性的性能。

关键词：行为识别；时空建模；3D CNN；时间迁移

Collaborative spatial-temporal networks for action recognition

ZHENG Yang, ZHANG Xudong
（School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China）

【Abstract】 Action recognition is an important research direction in the field of computer vision. It is widely used in video surveillance, crowd analysis, human-computer interaction, virtual reality and other fields. Spatial-temporal modeling is an important part of video action recognition, which could greatly improve the accuracy of behavior recognition. 3D CNNs can learn powerful spatial-temporal representations but are computationally intensive, which make them expensive to deploy; Improved 2D CNNs with a temporal shift can efficiently perform temporal modeling by shifting the feature map along the temporal dimension. However, temporal shift does not allow for the adaptively reweighting spatial-temporal features. Previous works have not explored the combination of the two types of methods to better modeling spatial-temporal information. This paper proposes a collaborative network that effectively combines a 3D CNN and 2D temporal shift. In particular, a new collaborative spatial-temporal module (CSTM) is introduced to learn spatial-temporal features jointly to integrate attention mechanism. And the paper verifies the effectiveness of CSTM on temporal-related datasets (i.e., Something-Something v1, v2, Jester) and obtains superior results compared to the existing state-of-the-art methods.

【Key words】 action recognition; spatial-temporal modeling; 3D CNN; temporal shift

0 引 言

行为识别旨在识别视频中的人类动作。以前的行为识别方法[1-2]只使用静态信息就获得良好的结果，这通常通过从稀疏采样帧中观察状态变化以推断出动作类别。然而，现实生活中的视频数据包含时序信息。因此，时空特征对行为识别具有重要意义。时空特征编码了不同时间空间特征之间的关系[1,3-4]。

基于手工制作的行为识别方法被开发已久，主要包括几种不同的时空特征检测器。其中，The Improved Trajectories[2]被认为是当下最有效的传统算法，能沿着光流引导的轨迹提取局部特征。然而，低级别的手工制作的特征对细粒度的动作类别缺乏很强的代表性和识别能力。

卷积神经网络（Convolutional neural networks，CNNs）在行为识别[1-2,4]上取得了相当大的成功。与传统的方法相比，这些方法的识别精度提升很大[5-6]。在这些方法中，Sudhakaran 等人[7]提出了一种 Gate-shift 模块（GSM），该模块使用分组门控和时间移位来学习时空表示。时间移位在文献[8]中提出，就是通过在信道维度上移动特征，以实现相邻视频帧间的特征交互。此外，时间位移模块可以区分具有相似外观的动作类别[7]。在以前的方法中，时空特征通常是使用单一类型的卷积来捕获的，即

基金项目：国家自然科学基金（61876057，61971177）。

作者简介：郑 阳（1996-），男，硕士研究生，主要研究方向：计算机视觉；张旭东（1966-），男，博士，教授，主要研究方向：计算机视觉、模式识别。
通讯作者：郑 阳 Email：zhengyangjuly@163.com

收稿日期：2021-03-11

vanilla 2D CNN、3D CNN、或改进的具有时间位移操作的 2D CNN。这意味着以前的方法不能做出依赖于数据的决策、进而有选择地使特征通过不同的卷积结构。本次研究的目的是基于时空特征设计不同的卷积操作。例如,某些特征可以更好地用 3D 特征提取,另一些可以更好地用改进的 2D CNN 表示。

在本文中,提出了一种新的协同时空特征学习模块(CSTM),其中结合了 3D CNN 和时间位移操作来共同获得空间和时间特征。时间位移分支通过沿时间维度移动信道与相邻帧交换信息。3D CNN 分支基于滑动窗口[9]对输入视频的短期时间上下文进行建模。基于 SENet[10]中,提出了一种协同注意力机制用以有效地融合来自 3D CNN 和时间位移分支的特征。该算法可以增强重要特征并减弱无关特征。和之前方法不同的是,本文的方法可以更有效地学习时空信息,从而自动地学习一种选择策略。3D CNN 分支学习动作状态变化较大的动作类别;时间位移分支区分具有相似外观的动作类别。

总结本文的贡献如下:

(1)文中将 3D CNN 和时间位移操作结合起来,在一个统一的框架中编码互补的时空特征。

(2)文中提出了一种新的具有嵌入式注意机制的协作时空学习模块,用以融合从 3D CNN 和时间位移获得的响应。

(3)与现有的最先进的方法相比,本文的方法在几个时间相关的数据集上取得了竞争性的性能,包括一些 Something-Something v1、v2、Jester。

1 相关工作

(1)2D CNN。在之前的工作中,2D CNNs 被用于视频动作识别,并取得了较好的效果[5-6,11]。Simonyan 等人[6]首先提出了一个针对 RGB 输入(空间流)和一个光流输入(时间流)的双流 CNN。Wang 等人[5]提出了一种针对双流结构的稀疏时间采样策略,并通过加权平均融合了 2 个流(temporal segment network;TSN)。Feichtenhofer 等人[12]探索了 2 个流的融合方法来学习时空特征。尽管上述方法是高效和轻量级的,却只是使用加权平均或平均池化来融合特征,而忽略了时间顺序或更复杂的时间关系。为了克服这个缺点,Zhou 等人[13]提出了一个时间关系网络(TRN)来学习视频帧之间的时间依赖关系。Wang 等人[14]提出了一种 non-local 神经网络来建模远程依赖。Lin 等人[8]提出了一种基于 TSN 的时间位移模块,可通过沿时间维度移动特征通道用以进行时空建模。这些方法都是基于 2D CNN+后融合,且被认为是建模时空关系的有效方法。

(2)3D CNN。学习帧间时空特征的另一种方法是使用 3D CNN[1,4]。Tran 等人[1]使用 3D 卷积(C3D)从一序列密集帧中提取时空特征。Tran 等人[4]进一步将 3D 卷积引入 ResNet 结构,对 C3D 进行了改进。SlowFast[12]包括捕获空间语义的慢路径和以一个细粒的时间分辨率捕获运动信息的快速路径。然而,3D CNN 包含了大量的参数,也很难在现实世界加以部署。因此,本文的工作只在特定的几个网络层使用三维卷积来学习时空信息。这将使计算量最小化,同时也确保了较高的效率。

(3)2D CNN+3D CNN。有几项工作已经研究了有效性和计算成本之间的权衡。Zolfaghari 等人[9]在一个 2D 的时间融合网络后,增加了一个 3D 残差网络。Luo 等人[15]提出了利用 2D 和 3D 卷积的时空交互建模方法。这些方法的性能有所提升,同时减少了参数的数量。文中的模型是基于混合 2D 和 3D CNN,即使用 2D 和 3D CNN 同时提取时空信息。特别是,文中的网络包含了基于数据的决策策略,即根据特征选择不同的卷积结构。此外,文中的模型只使用 RGB 稀疏帧作为输入,而不是 RGB 帧和光流的组合。

2 方法

受启发于 Gate-Shift 模块(GSM)[7],本文提出一种创新的协同时空模块(CSTM)。在本节,首先描述 GSM 的构成细节。之后,将详解剖析本次研发设计的模块。

2.1 Gate-Shift 模块

Gate-Shift 模块融合了 GST[15]和 TSM[8]用以构建高效的时空特征提取器。图 1(a)展示了 GSM 的概念图。其中具备一个学习的空间门控单元。这个门控单元通过时间迁移操作选择性地通过部分特征信息。图 1(b)阐述了 GSM 详细的组成结构。具体由一个分组门控单元和一个前后时间迁移操作组成。其中,分组门控单元是用一个三维卷积和一个 tanh 激活层实现。因为文中的网络结构是基于 GSM 做进一步改进,所以详细描述 GSM 有助读者能全面了解文中的网络结构。文中以向量 $\boldsymbol{X}$ 表示 GSM 的输入,大小为 $C \times T \times H \times W$。这里 C 表示通道数量,W、H、T 分别表示特征图的宽、高和时间维度。$\boldsymbol{X}$ 沿着时间维度被分为 2 组,即 $[X_1, X_2]$,W 包

含 2 个 $C/2 \times 3 \times 3 \times 3$ 大小的门控核。$[Z_1, Z_2]$ 表示输出。整个模块的计算过程如下：

$$Y_1 = \tanh(W_1 * X_1) \odot X_1 \quad (1)$$

$$Y_2 = \tanh(W_2 * X_2) \odot X_2 \quad (2)$$

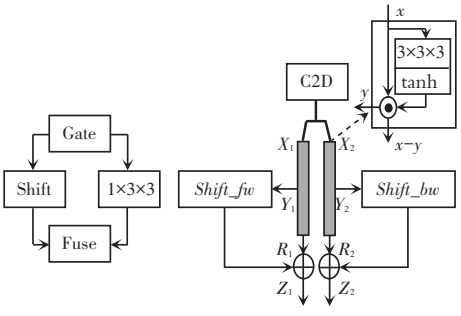
$$R_1 = X_1 - Y_1 \quad (3)$$

$$R_2 = X_2 - Y_2 \quad (4)$$

$$Z_1 = shift\_fw(Y_1) + R_1 \quad (5)$$

$$Z_2 = shift\_bw(Y_2) + R_2 \quad (6)$$

其中，‘$*$’表示卷积；‘$\odot$’表示 Hadamard 点乘；*shift_fw* 和 *shift_bw* 分别表示前向和后向的位移操作。公式(3)和公式(5)也可以分别表示为 $Z_2 = shift\_bw(Y_2) + R_2$ 和 $Z_1 = X_1 + R_1$，这种表示方法和 ResNet[11] 中的残差结构类似。



（a）GSM 模块概念图　　（b）GSM 详细结构

图 1　GSM 结构

Fig. 1　Structure of GSM

2.2　协同时空模块

本文提出的协同时空模块(CSTM)的结构如图 2所示。在协同时空模块中，为了进行交互，3D CNN 支路和时间迁移支路交叉加权彼此的中间特征。首先，2D 卷积对网络初始输入进行处理得到时空特征。然后得到的特征输出通过分组门控单元得到门控特征。门控特征随即被分别传递到 3D CNN 支路和时间迁移支路。其中，时间迁移支路是来自于 GSM，将沿着时序维度位移一部分特征图。

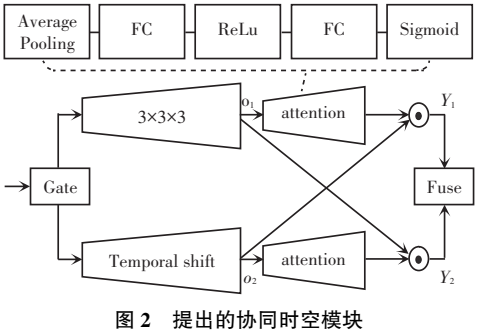


图 2　提出的协同时空模块

Fig. 2　Proposed CSTM

基于 SENet[10]，文中设计一种协同通道注意力模块用以有效地融合 3D CNN 支路和时间迁移支路。这个模块由一个 3D 平均池化操作、2 个全连接层和最后连接的一个通道级的缩放操作。3D 平均池操作将全局空间信息压缩成信道描述符，以便于利用全局感受野中的上下文信息。2 个全连接层目的在于完全捕获通道间的依赖关系。通道注意力模块动态地对通道级的特征进行重新校准。并经常被用于需要鉴别细粒度特征的任务中。这里将在下面的篇幅中介绍通道注意力模块地工作原理。首先，3D CNN 支路的输出(o_1)、迁移支路的输出(o_2)将分别通过通道注意力模块。由上一步获得的特征将与 o_1 和 o_2 进行通道级的相乘。输出 $Y = [Y_1, Y_2]$ 的计算过程如下：

$$Y_1 = o_1 \odot \delta_2(W_2\delta_1 W_1(Pooling(o_2))) \quad (7)$$

$$Y_2 = o_2 \odot \delta_2(W_2\delta_1 W_1(Pooling(o_1))) \quad (8)$$

其中，‘$\odot$’表示 Hadamard 乘；δ_1 表示 ReLU 函数；δ_2 表示 sigmoid 函数；W_1, W_2 分别表示 2 个全连接层。

最后，2 个分支的特征响应被进一步连接并简化为更紧凑的表示。

本文设计的协同时空模块被插入到 BN-Inception 和 InceptionV3 骨干网络中，如图 3 所示。因为 Inception 单元的其它支路中空间卷积的卷积核尺寸都很大，这严重影响到本文网络对空间特征的学习能力。所以研究中仅仅将协同时空模块插入到 Inception 单元中含有最少卷积数量的支路。

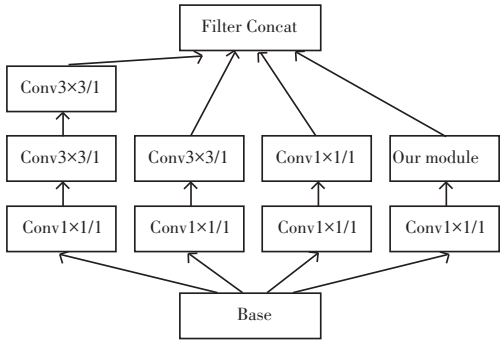


图 3　嵌入协同时空模块的 Inception 结构

Fig. 3　Inception with embedded CSTM

2.3　网络结构

整体的网络框架如图 4 所示。视频被分成 N 个相同大小的片段。从每个数据段中采样一个帧。文中采用 BN-Inception 和 InceptionV3 作为骨干网络。CSTM 随后插入到 Inception 单元中最少数量的卷积层分支中，以提取时空特征并进行时间融合。文中使用 TSN[5] 作为基础的框架结构，并且采用 BN-Inception[16] 和 InceptionV3[17] 作为文中的骨干网络。本章提出一种新的时空模块(Collaborative Spatial-Temporal Module，CSTM)进行视频中的时空建模。

文中提出的模块能和任意的 2D 卷积结合。后续的实验是将设计的模块插入到 BN－Inception 和 InceptionV3 中。

最终的预测结果是对每个帧的结果进行一种简单的平均池化操作。我们证明了在实验中采用的平均池化的融合方法的性能比与需要在网络高层上进行复杂融合的方法优越。原因是本文设计的模块在网络的中间层已经将时空特征进行了不断的融合。

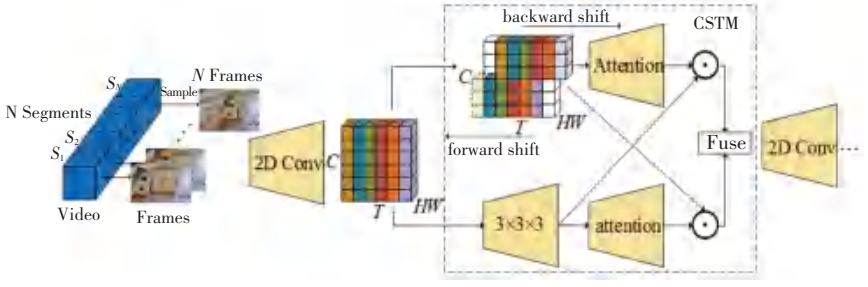


图 4 本文网络模型框架

Fig. 4 Architecture of the propsed network

3 实验和结果

3.1 数据集

研究中在 3 个公开的数据集上评估了本文的网络。对此拟做阐述分述如下。

(1) Something－Something v1[18], v2[19]。Something-Something 是人与物体交互的视频数据集。共包含 108 499 个视频,174 个类别。需要广泛的时间建模来区分这些细粒度类别。Something-Something v2 是第二个版本,其中含有220 847个视频,并且显著降低了标签噪声。

(2)Jester[20]。Jester 是手势识别的数据集。其中包含148 092个视频,27 个类别。

3.2 实验细节

实验中所用的工作站配置 CPU 为 Intel Xeon (R) E5-2620v2@2.1 GHz x 15,显卡为 2 × NVIDIA GTX2080Ti 12 G,内存为 128 G,系统为 Ubuntu 16.04LTS,使用编译软件为 Python 3.6,使用深度学习框架为 Pytorch[21]。整个网络使用随机梯度下降算法(Stochastic Gradient Descent,SGD)端到端进行训练。实验使用余弦学习率策略(cosine learning rate schedule),初始的学习率设置为 0.01。动量(*Momentum*)设置为 0.9,权重衰减(Weight Decay)设置为 0.000 5, *dropout* 设置为 0.25,批尺寸(Batch size)设置为 32。实验是在 Something－Something v1&v2 和 Jester 三个数据集上进行训练,训练的最大迭代次数为 60 个周期(epoch)。这里将采用 BN-Inception 和 InceptionV3 作为实验中的骨干网络,其输入图片的尺寸大小分别为 224 × 224 和 229 × 229。

训练使用交叉损失作为损失函数,如公式(9)所示:

$$L = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} y_{ji}\log \hat{y}_{ji} \tag{9}$$

其中, m 为批大小(batch_size);总的类别数为 n; 真实分布为 y_{ji}; 网络输出分布为 $\hat{y}_{ji}$。

3.3 对比实验

在本节中,在 Something-Something v2 数据集的验证集上进行了对比实验。为了验证文中模型的有效性,使用 BN-Inception 作为骨干网络,以 8 个视频帧作为网络的输入。下面将分别对协同时空模块的影响、注意力融合机制的有效性问题进行实验分析。最后,分别在 Something-Something v1 & v2 和 Jester 三个数据集上与当下先进的行为识别方法进行对比。研究过程详见如下。

3.3.1 协同时空模块的影响

在本小节,研究的目的是验证协同时空模块的有效性。首先,结合 3D CNN 和时间迁移模块。接下来进行实验以验证 3D CNN 能够学习到互补的时空特征信息。然而,实验中结合 3D CNN 和时间迁移模块的简单的加法融合方式并没有带来实验结果上的明显提升。文中将这种使用简单的加法融合方式的模型命名为 3D_shift_sum,并且以此作为基准。后续内容将进一步阐述了本文提出的融合策略有效地提升了模型性能。

研究展示了 3D CNN 在文中框架上的效果。这里将其与原始的 GSM 的结果对比见表 1。表 1 中,加粗表示最优性能。由表 1 可知,与 GSM 相比,3D_Shift_sum 和协同时空模块(CSTM)分别实现 0.42% 和 2.22% 的 top-1 准确率提升。这个结果表明了简单地对 3D CNN 和时间迁移模块(3D_Shift_sum)进

行融合只能带来微小的性能提升。而本文设计的协同时空模块展现出了显著的效果。协同时空模块使用交互的注意力机制用于融合 3D CNN 和时间迁移模块的特征信息。并能够很好地区分出细粒度的行为类别。研究中展示的一小部分行为类别的 top-1 准确率如图 5 所示。分析时注意到文中的模型在给定视频级的行为标签下能够学习到行为的状态变化。Something-Something v2 数据集中抽样帧如图 6 所示。在图 6 中，“pulling something from onto something”、“pulling something from right to left” and “pulling two ends of something so that it gets stretched”属于相同的粗粒度标签。然而，却在每个时间段展现出不同的状态变化。文中提出的协同时空模块足够有效地捕捉到了视频中行为的状态变化信息。

表 1　GSM，3D_Shift_sum 和 CSTM 在 Something-Something v2 验证集上的定量比较

Tab. 1　Performance comparison between GSM, 3D_Shift_sum and CSTM on the validation set of Something-Something v2

方法	Top-1	Top-5
GSM	61.23	87.29
3D_Shift_sum	61.65	87.56
CSTM	**63.45**	**88.79**

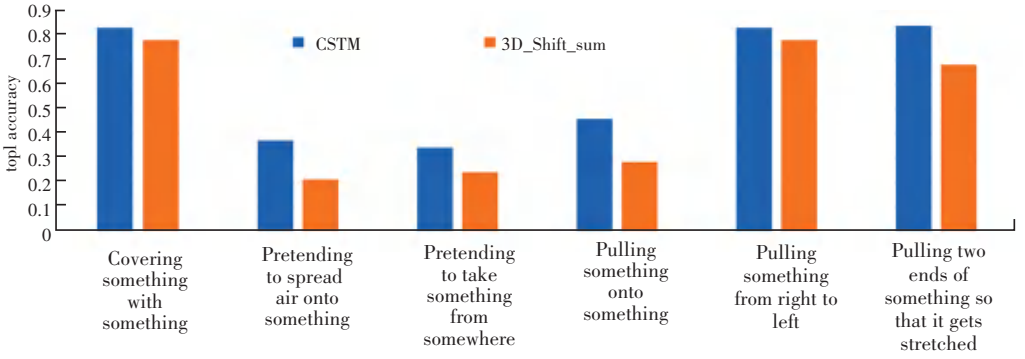


图 5　CSTM and 3D_Shift_sum 中 top-1 准确率提升较大的行为类别比较

Fig. 5　Comparison of categories with greater top-1 accuracy improvements between CSTM and 3D_Shift_sum

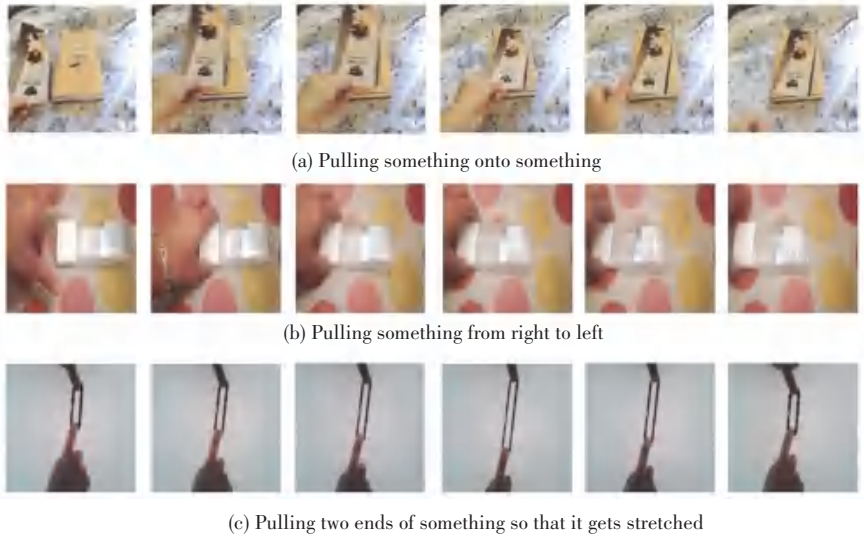


图 6　Something-Something v2 数据集上抽样帧

Fig. 6　Sampled frames from the Something-Something v2 dataset

图 7 展示了 3D_Shift_sum 和 CSTM 模型下 15 个类别的 t-SNE 可视化结果。由图 7 可以得出结论：本文的网络可以区分出细粒度的行为标签，比如“Poking a stack of something without the stack collapsing”、“Trying to pour something into something, but missing so it spills next to it” and “Pulling two ends of something so that it gets stretched”。而且，本文的网络可以学习到状态变化。

另外，“Taking something from somewhere”、“Moving part of something” and “Pulling something out of something”等行为类别的准确率只有大约 20%。这些表现较差的行为类别呈现出行为在视频中的持续时间较短和变化较为缓慢的共同特点。本文的方法在这些行为类别上表现出短板。因此，设计一个更加细粒度的网络用以挖掘时空特征信息是未来的研究方向。

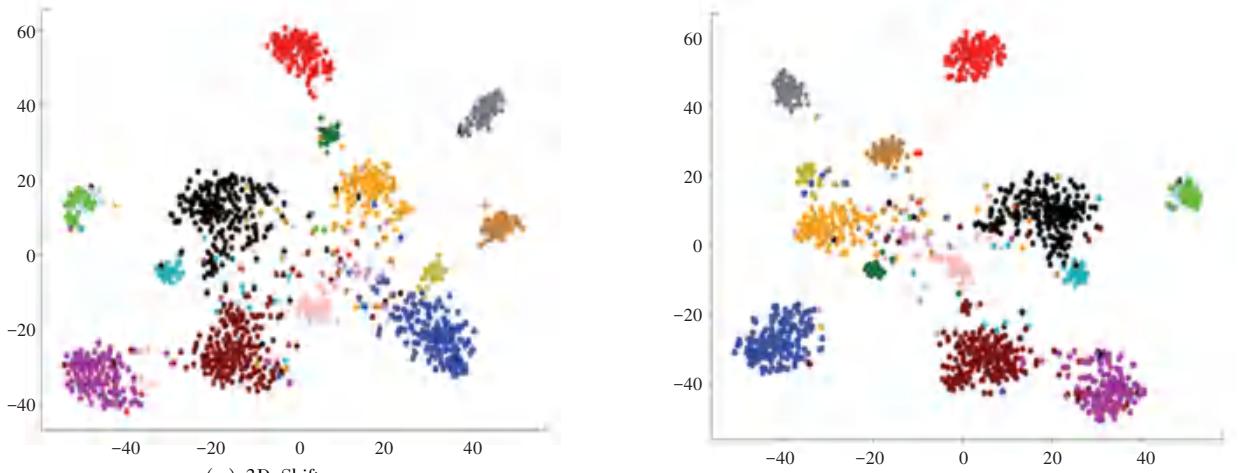


图 7 3D_Shift_sum 和 CSTM 模型下 15 个类别的 t-SNE 可视化结果

Fig. 7 t-SNE plot of the video samples of the 15 classes using the deep features from 3D_Shift_sum and CSTM

3.3.2 注意力融合机制的有效性

在前面融合 3D CNN 和时间迁移模块用以提取特征信息之后,文中提出了一种基于通道注意力机制的融合策略。为了验证本文提出的融合策略对所提出的模型来说是最合适的,本文结合通道注意力机制与 GSM,并且对比了其与本文设计的协同时空模块。结果展示见表 2。表 2 中,加粗表示最优性能。协同时空模块实现了 1.32%的 top-1 准确率提升;这显然清楚表明了协同时空模块的优越性能。注意力机制与原始的 GSM 结合时并没有取得最优的结果;相反,将其与本文的网络结合时实现了比较优越的结果。

3.3.3 与先进方法的对比

本文在 Something-Something v1&v2 和 Jester 三个数据集上与当下先进的方法的行为识别算法进行了对比。表 3 是在这三个数据集上的定量结果。为了公平起见,仅考虑在 RGB 输入下的结果。表 3 中,加粗表示最优性能。

表 2 GSM+attention 和 CSTM 在 Something-Something v2 验证集上的定量比较

Tab. 2 Performance comparison between GSM + attention and CSTM on the validation set of Something-Something v2

方法	Top-1	Top-5
GSM+attention	61.23	87.65
CSTM	**63.45**	**88.79**

表 3 在 Something-Something v1, v2, Jester 数据集上与先进方法的对比

Tab. 3 Comparison of CSTM with state-of-the-art methods on Something-Something v1, v2, Jester

方法	骨架网络	预训练数据集	帧数	Something v1		Something v2		Jester	
				Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
TSN	ResNet-50	Kinetics	8	19.70	46.60	27.80	57.60	81.80	99.00
TRN multiscale	BN_Inception	ImageNet	8	34.44	-	48.80	77.64	95.30	-
TSM	ResNet-50	Kinetics	8	43.40	73.20	58.20	84.80	94.40	99.70
	ResNet-50	Kinetics	16	44.80	74.50	58.70	84.80	-	-
GSM	BN_Inception	ImageNet	8	47.24	-	61.23	87.29	95.73	99.83
S3D	BN_Inception	ImageNet	64	47.30	78.10	-	-	-	-
S3D-G	BN_Inception	ImageNet	64	48.20	78.70	-	-	-	-
STM	ResNet-50	ImageNet	8	-	-	-	-	96.60	**99.90**
CSTM	BN_Inception	ImageNet	8	47.96	76.77	63.45	88.79	96.40	99.84
	Incepiton v3	ImageNet	8	49.24	77.56	63.95	**89.11**	96.51	99.86
	BN_Inception	ImageNet	16	49.84	78.71	63.54	88.99	96.48	99.87
	Incepiton v3	ImageNet	16	**51.27**	**79.90**	**64.04**	88.54	**96.61**	99.86

在 Something-Something v1&v2 数据集上，本文的模型在仅有 8 个帧输入的情况下比绝大多数先进的方法都要优越。本文的方法优于后期融合方法 TSN 和 TRN，因为能更好地编码空间和时间特征。在 Something-Something v1 数据集上，本文的模型在使用更少的帧的情况下表现出比 S3D[3] 更好的结果。在 Something-Something v2 数据集上，本文的模型在 TSN 基础上则又提升了 35.65% top-1 准确率和31.19% top-5 准确率。尽管本文的模型仅仅使用 RGB 视频帧作为输入，但是获得了优越的结果。

4　结束语

在本文中，提出了一种有效的用于行为识别任务的网络，称之为协同时空模块（CSTM）。设计上有效地结合了三维卷积和时间迁移模块，并且可以互补地学习视频数据中的时空特征。实验中在几个与时间相关的数据集（Something-Something v1 & v2 和 Jester）上评估了本文提出的网络，均展现了竞争性的性能。此外，本文设计的网络模型在仅使用 RGB 输入的情况下获得了与现有先进方法相比更好的结果。

参考文献

[1] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile:IEEE ,2015: 4489-4497.

[2] WANG Heng, SCHMID C. Action recognition with improved trajectories[C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney, NSW, Australia :IEEE, 2013: 3551-3558.

[3] XIE Saining , SUN Chen, Huang J , et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[J]. arXiv preprint arXiv:1712.04851,2017.

[4] TRAN D, RAY J, SHOU Z, et al.Convnet architecture search for spatiotemporal feature learning[J]. arXiv preprint arXiv: 1708.05038, 2017.

[5] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition[M]// LEIBE B, MATAS J, SEBE N, et al. Computer Vision - ECCV 2016. Lecture Notes in Computer Science. Cham:Springer, 2016,9912: 20-36.

[6] SIMONYAN K, ZISSERMAN A. Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Advances in Neural Information Processing Systems, 2014, 1.

[7] SUDHAKARAN S, ESCALERA S, LANZ O. Gate-shift networks for video action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington:IEEE,2020: 1102-1111.

[8] LIN J, GAN C, HAN S. Temporal shift module for efficient video understanding[J]. CoRR abs/1811.08383 ,2018.

[9] ZOLFAGHARI M, SINGH K, BROX T, etal. Eco: Efficient convolutional network for online video understanding[C]// Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany:Springer Science+Business Media, 2018 :695-712.

[10] HU Jie, LI Shen, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8):2011-2023.

[11] CHRISTOPH R, PINZ F A. Spatiotemporal residual networks for video action recognition[C]//Advances in Neural Information Processing Systems. London,England:The MIT Press, 2016: 3468-3476.

[12] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea:IEEE,2019:6202-6211.

[13] ZHOU B , ANDONIAN A , OLIVA A , et al. Temporal relational reasoning in videos[M]// FERRARI V, HEBERT M, SMINCHISESCU C, et al. Computer Vision-ECCV 2018. Lecture Notes in Computer Science. Cham:Springer, Cham,2018,11205: 831-846.

[14] WANG X , GIRSHICK R , GUPTA A , et al. Non-local Neural Networks[J]. arXiv preprint arXiv:1711.07971,2017.

[15] LUO C, YUILLE A L. Grouped spatial-temporal aggregation for efficient action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea:IEEE,2019: 5512-5521.

[16] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// International Conference on Machine Learning. Miami, Florida, USA :PMLR, 2015: 448-456.

[17] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. LasVegas,NV, USA:IEEE,2016: 2818-2826.

[18] GOYAL R , KAHOU S E , MICHALSKI V , et al. The "somethingsomething" video database for learning and evaluating visual common sense[C]// 2017 IEEE International Conference on Computer Vision (ICCV).Venice, Italy:IEEE, 2017:5843-5851.

[19] MAHDISOLTANI F, BERGER G, GHARBIEH W, et al. Fine-grained video classification and captioning[J]. arXiv preprint arXiv:1804.09235, 2018.

[20] MATERZYNSKA J , BERGER G , BAX I , et al. The Jester dataset: A large-scale video dataset of human gestures[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW).Seoul, Korea (South) :IEEE, 2019:2874-2882.

[21] PASZKE A, GROSS S, MASSA F, et al.Pytorch: An imperative style, high-performance deep learning library[J]. arXiv preprint arXiv:1912.01703,2019.

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