Computer Science and Application 计算机科学与应用, 2021, 11(6), 1607-1616
Published Online June 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.116166

Hans汉斯

基于注意力机制与双向LSTM的行为识别

张玉铭，吴克伟，金依珂，周龙辉

合肥工业大学计算机与信息学院，安徽 合肥
Email: 1083546770@qq.com

收稿日期：2021年5月1日；录用日期：2021年5月26日；发布日期：2021年6月2日

摘 要

采用光流作为运动特征进行行为识别需要预先计算并存储光流，需要巨大的计算成本和存储资源，并且由于光流特征主要表征了相邻帧之间的运动特征，导致行为识别中存在长依赖问题。针对这些问题，本文提出了一种新的运动特征建模方式以取代光流特征，并且提出了一种长依赖时序运动建模模块。实验结果表明，本文提出的方法在增加极低的计算成本的情况下，能更好的对远距离图像帧间的时序上下文信息建模，显著提高行为识别的准确度。

关键词

行为识别，光流，运动特征，长依赖问题，时序上下文信息

Action Recognition Based on Attention and Bi-LSTM

Yuming Zhang, Kewei Wu, Yike Jin, Longhui Zhou

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui
Email: 1083546770@qq.com

Received: May 1st, 2021; accepted: May 26th, 2021; published: Jun. 2nd, 2021

Abstract

Using optical flow as motion features for action recognition requires pre-computation and storage of optical flow, which requires huge computational cost and storage resources. And optical flow features mainly characterize the motion features between adjacent frames, which leads to long-dependency problems in action recognition. To address these problems, this paper proposes a new way of modeling motion features to replace optical flow features and proposes a long-dependency temporal motion modeling module. Experimental results show that the proposed me-

文章引用: 张玉铭, 吴克伟, 金依珂, 周龙辉. 基于注意力机制与双向 LSTM 的行为识别[J]. 计算机科学与应用, 2021, 11(6): 1607-1616. DOI: 10.12677/csa.2021.116166

thod in this paper can better model the temporal context information between long-range frames and significantly improve the accuracy of action recognition with very low increase in computational cost.

Keywords

Action Recognition, Optical Flow, Motion Features, Long-Dependency Problems, Temporal Context Information

1. 引言

行为识别是计算机视觉领域非常有挑战性的课题，研究的是视频中目标的行为动作，在视频监控、人机交互等领域有着重要的应用。在行为识别任务中，不仅要分析图像中目标的空间信息，还要提取时间维度上的时序特征。近几年，针对行为识别的研究主要围绕双流法和 3D 卷积进行展开。双流法包括两个并行分支：RGB 空间流与时序流。在时序流中，通常采用光流作为运动特征，这也导致双流法需要预先计算并存储光流。因此，双流法需要巨大的计算和存储资源。3D 卷积在 2D 空间卷积的基础上增加时间维度，学习时序特征，维度的增加也导致了计算量的增长。此外，双流法与 3D 卷积只能对局部的时间段内的时序特征进行提取，对长时视频中的长依赖问题存在较大的局限性。

针对以上存在的问题，我们提出了两个模块：通道级运动特征编码模块(CME)与长依赖时序运动建模模块(LDTM)。CME 模块采用帧差法的思想，对相邻帧之间的运动特征进行编码，取代光流特征。利用注意力机制在通道级对运动特征通道进行激励，对背景通道进行抑制。LDTM 模块利用双向 LSTM 网络，在增加较低的计算成本代价下，对远距离图像帧之间进行时序上下文关系建模，对解决行为识别中的长依赖问题提供了一种解决思路。

2. 相关工作

传统的行为识别方法一般是通过人工观察和设计，手动设计出能够表征动作特征的特征提取方法，比如方向梯度直方图(HOG)、光流直方图(HOF)、光流梯度直方图(MBH)、轨迹特征(Trajectories)以及人体骨骼特征等。在早期研究工作中，iDT (improved Dense Trajectories) [1]算法是性能最好的算法之一。近几年，随着深度学习的兴起与快速发展，越来越多基于深度学习的方法被提出应用于行为识别，例如比较经典的双流网络架构以及近几年兴起的 3D 卷积网络及其变种。

2.1. 双流网络架构

2014 年，Karen Simonyan 等学者提出了双流(Two-Stream) [2]卷积网络，利用帧图像和光流图像作为 CNN 的输入，将行为识别中的特征提取分为两个分支，一个是 RGB 图像分支提取空间特征，另一个是光流图像分支提取时间上的光流特征，最后结合两种特征进行行为识别。在此之后，许多基于双流架构的先进算法陆续被提出，比如 TSN、TRN [3] [4]等。TSN 解决了传统双流法无法解决的长视频问题，提出将长视频切割成等长的若干视频段，每个视频段随机选取一帧作为输入提取特征，最后再进行特征融合。双流法的性能超越了传统手工特征提取方法，但是由于双流法需要预先计算光流提取光流特征，因

此带来巨大的计算成本和存储需求，于是许多学者开始探寻不需要计算光流的方法，典型的方法例如 3D 卷积。

2.2. 3D 卷积网络

C3D [5]作为 3D 卷积的早期尝试，将 VGG16 模型中的 2D 卷积全都换成 3D 卷积。C3D 虽然极具开创性，但是并没取得令人满意的效果，I3D [6]提出将 2D 卷积训练权重在时间维度重复，用 ImageNet 预训练的权重初始化 3D 卷积，大幅提升了 3D 卷积网络的性能。3D 卷积是 2D 卷积的扩展，虽然取得了较大的性能提升，证明了 3D 卷积网络建模时序特征的可行性，但是 3D 网络架构相比于 2D 卷积网络多了时间维度，因此也带来了巨大的计算量。为了降低 3D 卷积的计算量，许多工作对 3D 卷积操作进行改进以实现更少的参数量和更高的性能。P3D [7]提出将 3 × 3 × 3 的 3D 卷积拆分成 1 × 3 × 3 的 2D 空间卷积和 3 × 1 × 1 的 1D 时序卷积，前者建模空间特征，后者建模时序特征，并采用不同的排列方式和残差连接，设计出多种变形结构。ECO [8]在网络浅层采用 2D 卷积编码图像帧，在网络深层再用 3D 卷积对拼接起来的特征图进行编码。X3D [9]从 2D 卷积的架构延伸，选择某个需要扩张的维度并固定住其他维度，将 2D 卷积核从不同维度扩展成 3D 卷积核。此外，X3D 采用了通道可分离卷积网络，从而得到一个轻量级的网络模型。

除了双流模型与 3D 卷积以外，也有一些其他的工作聚焦于解决行为识别中的计算资源巨大、长距离依赖建模困难等问题。Chao-YuanWu 等学者发现多数研究直接使用视频作为模型输入，从而导致数据量过大，因此他们参考视频压缩的技术，对视频进行压缩[10]。为了避免光流的计算和储存，Hidden Two-Stream [11]让模型先产生类似光流特征，再参考双流网络架构模型，分开处理时序特征和 RGB 特征。TSM [12]提出在 2D 卷积中沿时间轴交替拼接特征通道，这个操作不需要任何额外计算成本和参数，解决了 3D 卷积网络计算成本高的问题，同时也保证了时序特征的连续性。

此外，一些学者提出用新的运动特征编码方式来取代光流特征，降低计算成本。Boyuan Jiang 等人提出了一个 STM [13]模块，它包含一个通道级的时空模块(CSTM)来呈现时空特征和一个通道级的运动模块(CMM)来有效地编码运动特征。Yan Li 等人认为时序建模是行为识别的关键，通常考虑短程运动和远程聚集，提出了一个时序激发和聚集模块(TEA) [14]，包括一个运动激发(ME)模块和一个多时序聚集(MTA)模块，专门用于捕获短距离和长距离的时序运动特征。除了取代光流特征，也有学者开始将研究方向投向更加轻量级的模型架构设计，例如采用 2D + 1D [15]卷积、shifting 或注意力机制等等。

2.3. 长短期记忆网络

在行为识别任务中，输入的视频段可以看作是若干视频帧按一定时序进行排列得到，当我们的模型在学习某一帧视频帧特征信息时，我们需要模型将前面视频帧所学的特征信息与当前帧联系起来，即学习视频中的上下文信息。传统的神经网络无法做到这一点，循环神经网络(RNNs)解决了这一点，循环神经网络带有自循环结构，使得特征信息可以持久保存。但是当视频帧增多，视频时段较长时，我们需要学习更远距离的上下文，循环神经网络的效果就不是很明显了。长短期记忆网络(LSTM)便是针对视频理解、自然语言处理及语音识别等时序任务中存在的长依赖问题而提出的一种网络模型。长短期记忆网络是一种特殊的循环神经网络，由 Hochreiter 和 Schmidhuber (1997)提出，它能够学习视频、语音或文本等数据中的长距离依赖关系。在随后的发展中，许多研究人员针对计算机视觉任务中出现的不同问题，对 LSTM 进行改进和推广，出现很多 LSTM 变种，例如双向 LSTM、GRU 等，它们在许多问题上起到很好的效果，得到了广泛使用。

在本文中，我们的主要贡献总结如下：

(1) 提出了一个基于注意力机制的通道级运动特征编码模块(CME)，学习相邻帧间的时序上下文，在小幅增加计算成本的情况下显著提高模型性能；

(2) 提出了一个长依赖时序运动建模模块(LDTM)，对行为识别中远距离帧间的上下文关系建模，有效降低了长依赖问题对行为识别的影响。

3. 方法论

3.1. 整体网络模型架构

在这一节中，我们提出了两个模块(通道级运动特征编码模块和长距离时序运动特征建模模块)来提取视频运动特征，其中，通道级运动特征编码模块提取视频段相邻帧间短时运动特征，长依赖时序运动建模模块(LDTM)提取长距离时序运动特征。

如图 1，展示了我们的整体网络模型结构图。我们选用 ResNet50 作为卷积网络，并用 CME Block 替换 ResNet50 中每一个卷积层的 Block，用以提取相邻帧间的运动特征；此外，我们在 Conv Layer2、Conv Layer3 及 Conv Layer4 后引出一个并行分支接入长依赖时序运动建模模块(LDTM)，并与每一个卷积层的输出进行融合，作为后续卷积或者全连接网络的输入。在接下来的小节中，我们将详细介绍我们提出的通道级运动特征编码模块(CME)以及长依赖时序运动建模模块(LDTM)。

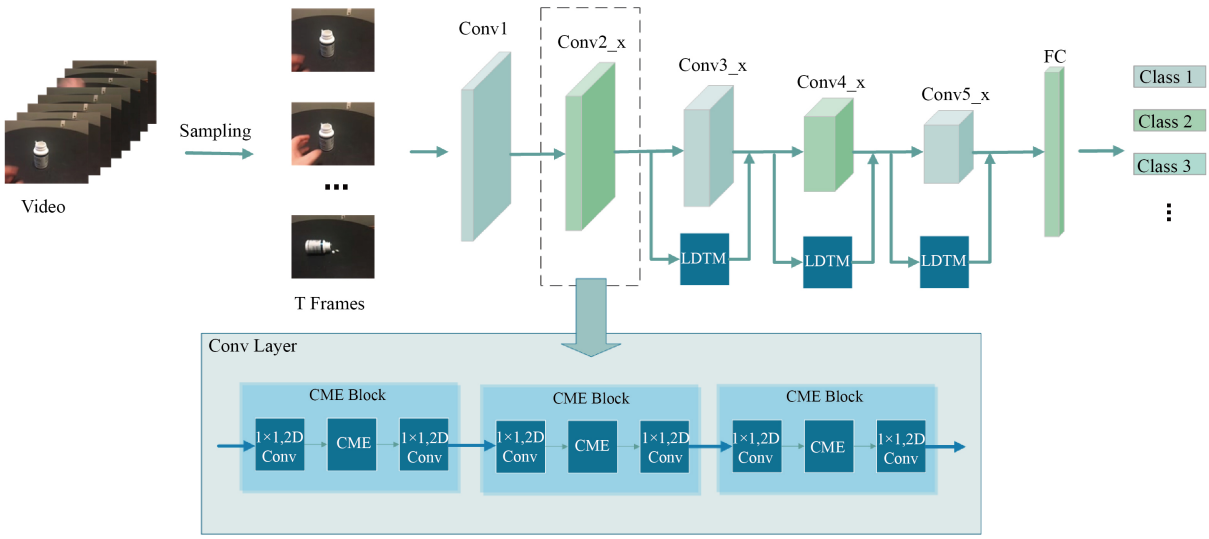


Figure 1. Overall structure of our methods
图 1. 整体网络模型结构图

3.2. 通道级运动特征编码模块(CME)

在之前的大量工作中，通常是通过提取光流特征来作为运动特征进行行为识别，但是计算光流特征需要巨大的计算资源和存储资源，效率比较低。因此，我们借鉴了帧差法检测运动目标的思想，提出了一个通道级运动特征编码模块 CME (Channel-wise Motion Encoding)来提取运动特征并激励运动信道，以取代光流特征。

如图 2 所示，在 CME 模块中，我们首先将输入的视频帧经过一个 1 × 1 的 2D 空间卷积降低特征通道维度，假设放缩因子为 r，输入维度为[N, T, C, H, W]，经过卷积后得到输出维度为[N, T, C/r, H, W]的特征图。将得到的特征图沿时间维度 T 拆分成单独的每一帧，然后将相邻帧进行差分，用后一帧图像特征 X_{t+1} 减去前一帧 X_t。由于视频中前景目标的运动，导致两帧之间目标位置发生变化，直接作差可能会导

致特征混乱。因此，为了消除特征混淆，我们先对后一帧用 3x3 的 2D 卷积进行空间特征变换，然后再用得到的变换特征与前一帧做差，从而得到相邻两帧之间的运动表征。我们假设第 t 帧到第 t+1 帧的运动特征为 M_t，那么：

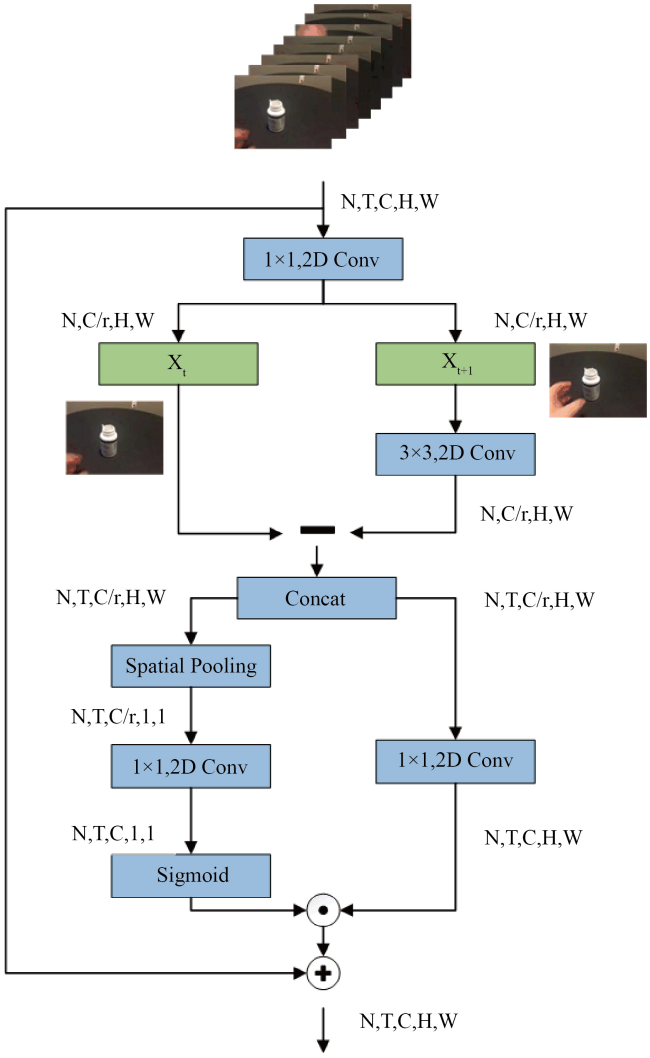


Figure 2. The structure of Channel-wise Motion Encoding (CME) module

图 2. 通道级运动特征编码模块(CME)结构图

$$M_t = Conv_{3\times3}\left(X_{t+1}\right) - X_t, \quad M_t \in R^{N\times C/r\times H\times W} \tag{1}$$

其中，X_t 表示第 t 帧图像特征，X_{t+1} 为第 $t+1$ 帧图像特征，$Conv_{3\times3}(\cdot)$ 表示 3 × 3 的 2D 空间卷积。将得到的相邻帧间运动特征拼接，得到输入视频的运动特征表示为 M，即：

$$M = Concat\left(M_1, M_2, \cdots, M_T\right), \quad M \in R^{N\times T\times C/r\times H\times W} \tag{2}$$

由于输入特征 X 经过了卷积通道降维，因此得到的视频运动特征 M 的维度为[N, T, C/r, H, W]。我们将得到的视频运动特征 M 输入两个并行分支，运动编码分支与运动信道激励分支。在运动编码分支，我们设计一个 1 × 1 的 2D 卷积对运动特征 M 进行编码，并将通道维度从 C/r 恢复到 C，确保特征信息不会损失，同时也便于与输入特征融合。假设运动编码后得到的特征为 M_{encode}，则有：

$$M_{encode} = Conv_{1\times1}(M),\quad M_{encode} \in R^{N\times T\times C\times H\times W} \tag{3}$$

其中，$Conv_{1\times1}(\cdot)$ 表示 1 × 1 的 2D 卷积操作。运动编码分支得到编码后的相邻帧之间的运动特征，但是特征通道包含了前景目标的运动特征以及背景特征。因此，我们采用注意力机制，对拼接后得到运动特征学习其通道的注意力权重，对于运动特征通道进行激励，对背景特征通道进行抑制。具体地，考虑到通道特征权重与空间位置无关，我们先将运动特征进行空间上的全局平均池化，得到空间无关的运动特征，记为 $M_{pooling}$，即：

$$M_{pooling} = Pooling(M),\quad M_{pooling} \in R^{N\times T\times C/r\times1\times1} \tag{4}$$

其中 $Pooling(\cdot)$ 表示空间池化操作，得到的 $M_{pooling}$ 维度为[*N*, *T*, *C*/*r*, *W*, *H*]。接下来，将 $M_{pooling}$ 经过卷积升维，并输入到 *sigmoid* 激活函数，得到注意力权重矩阵记为 *A*，即：

$$A = sigmoid\left(Conv_{1\times1}\left(M_{pooling}\right)\right),\quad A \in R^{N\times T\times C\times1\times1} \tag{5}$$

其中 $sigmoid(\cdot)$ 表示 *sigmoid* 激活函数，$Conv_{1\times1}(\cdot)$ 同上，表示 1 × 1 的 2D 卷积操作。将得到的通道注意力权重 *A* 与运动编码特征 M_{encode} 相乘，激励其中的运动特征通道，同时加入残差连接，保持原有输入特征不丢失，即：

$$Y = X + A\cdot M_{encode},\quad Y \in R^{N\times T\times C\times H\times W} \tag{6}$$

其中，*X* 为输入特征，*Y* 为模块的最终输出特征。显然，*X* 与 *Y* 具有相同的维度，即 CME 模块并不改变输入特征的输出维度，这允许我们可以将该模块嵌入模型的任何位置进行端对端的学习，在我们的实验中，我们将其嵌入 ResNet50 网络，并用以替换每一个卷积层 Block。

3.3. 长依赖时序运动建模(LDTM)

运动特征编码模块(CME)可以提取相邻帧之间的运动特征，激励运动特征通道，但对于时长较长的视频段时序运动特征提取存在局限性。因此，我们提出一个长依赖时序运动建模模块(Long-Dependency Temporal Motion)，如图 3 所示。

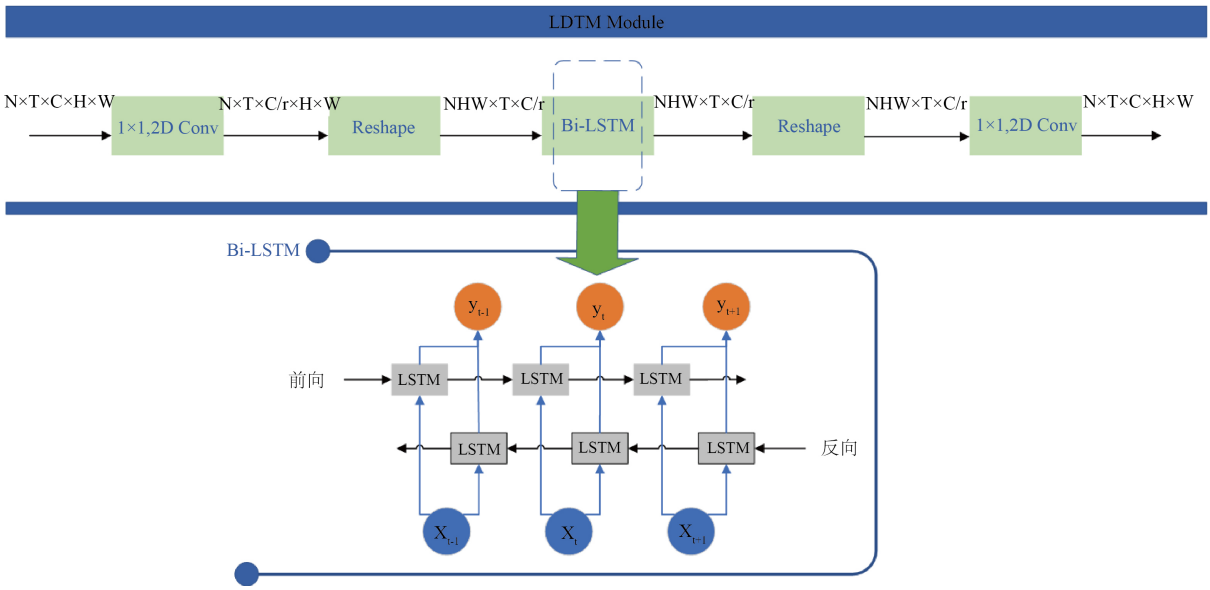


Figure 3. The structure of Long-Dependency Temporal Motion (LDTM) module
图 3. 长依赖时序运动建模(LDTM)模块结构图

首先，我们对输入的视频帧序列通过一个 1 × 1 的 2D 卷积对通道进行降维，缩放因子取 r = 16，然后将得到的特征经过维度变形转换后，输入一个双向 LSTM 网络(Bi-LSTM)。假设输入的视频帧序列为 $x \in R^{N\times T\times C\times H\times W}$，经 1 × 1 卷积后特征维度变为 $N\times T\times C/16\times H\times W$，然后通过 Reshape 操作将特征维度转换成 $NHW\times T\times C/16$。我们将若将 NHW 看作一个新的 Batchsize，则等价于将每一帧特征向量 $x_t \in R^{1\times C/r}$ 输入双向 LSTM 网络。在图 3 所示的结构图中，我们详细展示了双向 LSTM 模块的结构。具体地，我们将每一帧 $x_t \in R^{1\times C/r}$ 输入到一个前向 LSTM 网络和一个反向 LSTM 网络，将两者的输出特征进行融合得到每一帧的隐藏层输出为 $y_t \in R^{1\times C/r}$。显然，双向 LSTM 网络并不改变输入特征的输出维度。因此，我们将得到的双向 LSTM 输出特征进行逆操作(Reshape 和 1 × 1 卷积)后，并可将特征维度恢复与输入特征保持一致。我们将 LDTM 模块的操作流程表达如下：

$$y = Conv_{1\times1}\left(Rsh\left(BiLSTM\left(Rsh\left(Conv_{1\times1}(x)\right)\right)\right)\right), y \in R^{N\times T\times C\times H\times W} \tag{7}$$

其中，$Conv_{1\times1}(\cdot)$ 表示 1x1 的空间卷积，$Rsh(\cdot)$ 表示 Reshape 操作，$BiLSTM(\cdot)$ 表示双向 LSTM 的输出，x 与 y 分别代表输入及输出特征。

4. 实验结果与分析

4.1. 实验设置及数据集

实验设置。我们的实验选用 ResNet50 作为我们的基础卷积网络，并用我们提出的模块对 ResNet50 中的网络构件进行替换。为了便于与其他模型进行比较，我们在 Something-Something-v1 数据集上进行了实验。实验模型用 2 块 2080Ti 显卡进行训练，共训练 50 个 epoch。训练数据的 Batchsize 大小设定为 16，初始学习率设定为 0.005，并分别在第 30，40 以及 45 个 epoch 时将学习率缩小 10 倍。

数据集。Something-Something-v1 数据集是一个大型的带标签的视频段集合，这些视频段主要包括了人们对日常物品进行预定义的一些基本动作。该数据集包含 108499 个视频，并分为 174 个类别，其中训练集有 86017 个，验证集有 11522 个，测试集有 10960 个。

4.2. 与不同模型的性能对比

我们将我们的方法与当前最先进的一些行为识别模型在 Something-Something-v1 公开数据集上进行了实验对比。如表 1，可以发现，在相同的实验参数设置条件下，比如相同的帧采样、Backbone 网络以及预训练等，我们的方法相较于 TSN、TSM、GST 等优秀模型都有一定的性能提升。

Table 1. Comparison with state-of-the-art models on Something-Something-v1
表 1. 与当前先进的行为识别模型在 Something-Something-v1 数据集上进行对比

Method	Backbone	Frames × Crops × Clips	Flops	Param	Pre-train	Top-1 val (%)	Top-5 val (%)
TSN	ResNet50	8 × 1 × 1	33G × 1 × 1	N/A	ImageNet	19.70	46.60
TRN	BNInception	8 × 10 × N/A	16G × 10 × N/A	18.3M	ImageNet	34.40	63.20
TRN Two-Stream	BNInception	(8 + 8) × 10 × N/A	32G × 10 × N/A	36.6M	ImageNet	42.00	/
I3D	3D ResNet50	32 × 3 × 2	153G × 3 × 2	28.0M	ImageNet + Kinetics400	41.60	72.20
NL-I3D	3D ResNet50	32 × 3 × 2	168G × 3 × 2	35.3M	ImageNet + Kinetics400	44.40	76.00
NL-I3D&GCN	3D ResNet50	32 × 3 × 2	303G × 3 × 2	62.2M	ImageNet + Kinetics400	46.10	76.80

Continued

TSM	ResNet50	8 × 1 × 1	33G × 1 × 1	24.3M	ImageNet	45.60	74.20
TSM	ResNet50	16 × 1 × 1	65G × 1 × 1	24.3M	ImageNet	47.30	77.10
TSM&En	ResNet50	(8+16) × 1 × 1	98G × 1 × 1	48.6M	ImageNet	49.70	78.50
STM	ResNet50	8 × 3 × 10	33G × 3 × 10	N/A	ImageNet	49.20	79.30
STM	ResNet50	16 × 3 × 10	67G × 3 × 10	N/A	ImageNet	50.70	80.40
GST	ResNet50	16 × 1 × 1	59G × 1 × 1	N/A	ImageNet	48.60	77.90
GSM	BNInception	16 × 1 × 1	33G × 1 × 1	N/A	ImageNet	49.56	/
我们的方法	ResNet50	8 × 1 × 1	33G × 1 × 1	24.3M	ImageNet	47.89	76.78
	ResNet50	16 × 1 × 1	67G × 1 × 1	24.3M	ImageNet	49.74	79.03

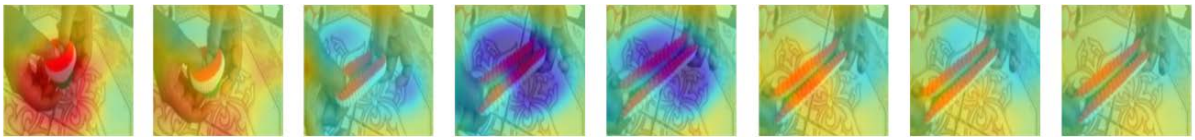
此外，我们也尝试了在 ResNet50 不同的层上加入 LDTM 模块，得到的对比实验数据如表 2。由于 ResNet50 网络的浅层特征图尺寸较大，语义信息较少，我们考虑跳过第一层卷积，而从第二个 Stage 开始输入 LDTM 模块。我们发现，在第 3 个卷积层(一个卷积层表示 ResNet50 网络的一个 Stage)后加入我们的 LDTM 模块，可以取得更好的效果。在较浅层或者更深层单独加入 LDTM 模块，由于浅层语义特征较少，而深层网络分辨率信息丢失，导致性能都有所下降。当然，在每一个 Stage 都加入 LDTM，能取得最好的效果，但也不可避免带来更大的计算量。

Table 2. Comparison after adding LDTM modules to different layers of ResNet50
表 2. 在 ResNet50 网络不同卷积层加入 LDTM 模块后的性能比较

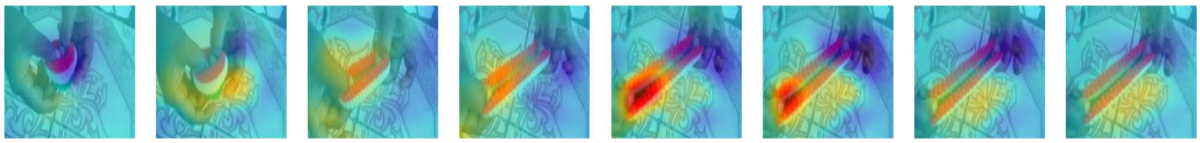
Stage	Top-1 val (%)	Top-5 val (%)
Conv2	42.6	71.8
Conv3	45.3	74.3
Conv4	44.6	73.4
Conv3 + Conv4	46.9	76.1
Conv2 + Conv3 + Conv4	47.4	76.2

4.3. 可视化分析

我们将我们的方法与 TSM 模型在 ResNet50 骨干网络第五层卷积 Conv5 后得到的特征图可视化。对比可以发现，我们的方法可以在长距离帧间进行关联的上下文学习。如图 4，在我们的方法中，第 1 帧与第 4、5 帧特征图中，人手在拉动物体时，前景目标区域的特征被分配了更大的权重和关注度，明显区别于背景目标。这也证明，我们的方法对长依赖时序运动特征建模是有效的。



(a)



(b)

Figure 4. The heatmap of "pulling two ends of something but nothing happens" action, where figure (a) shows our methods and figure (b) is TSM model.
图 4. "拉某物的两端但什么都没发生"行为视频段在 Conv5 后的特征热图，其中(a)表示我们的方法，(b) 表示 TSM 模型

5. 结束语

在行为识别任务中，短时相邻帧间采用光流特征作为运动特征进行行为识别存在着计算成本高、效率低等问题，而长时视频由于长依赖问题、远距离帧间上下文关系无法有效建模导致行为识别效果很差。针对以上的问题，我们提出了两个模块：通道级运动特征编码模块(CME)与长依赖时序运动建模模块(LDTM)。CME 模块在采用帧差法的思想，对相邻帧间的运动特征进行编码，并利用注意力机制在通道级对运动特征通道进行激励，对背景通道进行抑制，从而得到更加有效的短时运动特征。LDTM 模块利用双向 LSTM 网络，对远距离的帧间进行时序上的上下文关系学习，对解决行为识别中的长依赖问题提供了一种解决思路。CME 模块与 LDTM 模块均不改变输入特征的输出维度，这允许我们将模块嵌入到模型任何位置进行端对端的学习，在增加较低的计算成本代价下，显著增加了模型的识别性能。本文主要从短时运动特征编码与长时依赖问题两方面对行为识别进行思考，后续我们针对这两方面问题，将进一步优化提出的方法，以获得更好的性能。

参考文献

[1] Wang, H., *et al*. (2011) Action Recognition by Dense Trajectories. *The* 24*th IEEE Conference on Computer Vision and Pattern Recognition, CVPR* 2011, Colorado Springs, 20-25 June 2011, 3169-3176. https://doi.org/10.1109/CVPR.2011.5995407

[2] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. 28*th Annual Conference on Neural Information Processing Systems* (*NIPS* 2014), Montreal, 8-13 December 2014, 568-576.

[3] Wang, L., *et al*. (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition.

[4] Zhou, B., *et al*. (2018) Temporal Relational Reasoning in Videos.

[5] Tran, D., *et al*. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[6] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 4724-4733. https://doi.org/10.1109/CVPR.2017.502

[7] Qiu, Z.F., *et al*. (2017) Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 5534-5542. https://doi.org/10.1109/ICCV.2017.590

[8] Danelljan, M., *et al*. (2017) ECO: Efficient Convolution Operators for Tracking. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6931-6939. https://doi.org/10.1109/CVPR.2017.733

[9] Feichtenhofer, C. (2020) X3D: Expanding Architectures for Efficient Video Recognition. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 200-210. https://doi.org/10.1109/CVPR42600.2020.00028

[10] Wu, C.-Y., *et al*. (2018) Compressed Video Action Recognition. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6026-6035. https://doi.org/10.1109/CVPR.2018.00631

[11] Zhu, Y., *et al*. (2018) Hidden Two-Stream Convolutional Networks for Action Recognition. 14*th Asian Conference on*

Computer Vision, Perth, 2-6 December 2018, 363-378.

[12] Lin, J., *et al*. (2019) TSM: Temporal Shift Module for Efficient Video Understanding. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 7082-7092. https://doi.org/10.1109/ICCV.2019.00718

[13] Jiang, B.Y., *et al*. (2019) STM: SpatioTemporal and Motion Encoding for Action Recognition. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 2000-2009. https://doi.org/10.1109/ICCV.2019.00209

[14] Li, Y., *et al*. (2020) TEA: Temporal Excitation and Aggregation for Action Recognition. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 906-915. https://doi.org/10.1109/CVPR42600.2020.00099

[15] Tran, D., *et al*. (2018) A Closer Look at Spatiotemporal Convolutions for Action Recognition. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6450-6459. https://doi.org/10.1109/CVPR.2018.00675