基于时空信息辅助监督的语言-视频对比学习模型

张冰冰[1),2)] 张建新[1)] 李培华[2)]

[1)](大连民族大学计算机科学与工程学院 辽宁 大连 116650)

[2)](大连理工大学信息与通信工程学院 辽宁 大连 116033)

摘 要 同时使用语言和图像两种模态信息的神经网络模型在计算机视觉领域取得了很大进展.一些将其用于视频识别任务的工作,存在未考虑视频中丰富的时间-空间信息、用于描述类别的文本过于简单等不足.对此,本文提出了基于时空辅助信息监督的语言-视频对比学习模型.对于视频编码,提出了基于类别词元的时序加权位移模块进行时序建模,使得时序信息在网络从底层到高层的各个层次传播;而且还提出了时空信息辅助监督模块,深入挖掘视觉词元中蕴含的丰富时空信息.对于语言编码,提出了一种基于大语言模型的提示学习方法,对行为类别文本描述进行扩展,生成具有丰富上下文语义信息的文本描述.实验部分,本文提出的模型在4个视频行为识别数据集mini-Kinetics-200、Kinetics-400、UCF101和HMDB51上,达到了优于当前最先进方法或与当前最先进方法识别准确率相当的水平,比基线方法的识别准确率分别提升了2.5%、0.3%、0.6%和2.4%.

关键词 行为识别;多模态模型;时序建模;时空信息辅助监督;提示学习

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2024.01769

Contrastive Language-Video Learning Model Based on Spatio-Temporal Information Auxiliary Supervision

ZHANG Bing-Bing[1),2)] ZHANG Jian-Xin[1)] LI Pei-Hua[2)]

[1)](*School of Computer Science and Engineering, Dalian Minzu University, Dalian, Liaoning* 116650)

[2)](*School of Information and Communication Engineering, Dalian University of Technology, Dalian, Liaoning* 116033)

Abstract Video action recognition is one of the hot topics in the field of computer vision, which has attracted the attention of many researchers in recent decades. The basic method of video action recognition is widely used in Internet video audit, video surveillance, human-computer interaction and other fields. The main body of video is usually human. Because of the complexity and variability of human action categories and environment in real life, and the huge amount of video, it requires high computing devices, which brings great challenges to the task of video action recognition. In the field of video surveillance, most of the existing systems only record abnormal actions and cannot recognize it in real time, so they cannot realize real intelligence; while in the field of Internet video audit, a lot of manual audits is often needed, which can't recognize human action in real time. Video can usually be regarded as images that change with time. This special image data contains rich information. To recognize actions from video, it is not only necessary to

收稿日期:2023-05-23;在线发布日期:2024-04-25. 本课题得到国家自然科学基金(61971086、61972062)、吉林省科技厅科技发展计划项目(20230201111GX)和辽宁省应用基础研究计划项目(2023JH2/101300191、2023JH2/101300193)资助. **张冰冰**,博士,讲师,中国计算机学会(CCF)专业会员,主要研究领域为视频行为识别,图像分类和深度学习.E-mail:icyzhang@dlnu.edu.cn. **张建新**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为智能医学影像分析和图像/视频识别. **李培华**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为图像/视频识别、目标检测和语义分割.E-mail:peihuali@dlut.edu.cn.

obtain the spatial information of the image at each moment, but also to capture the temporal reasoning information between frames, and more importantly, to obtain the spatio-temporal information. To this end, researchers have developed many network architectures for video action recognition tasks, which can be divided into the following four categories: two-stream convolutional neural networks (CNNs) based methods, 3D CNNs based methods, 2D convolutional network with spatio-temporal modeling module, and visual Transformer-based networks. The use of Transformer-based network models that integrate both language and image modalities has made great progress in the field of computer vision. There are three representative research works in computer vision tasks related to images: namely Contrastive Language-Image Pre-training (CLIP) model, A Large-scale Image and Noisy-text embedding (ALIGN) model and Florence model. However, when these models are applied to video recognition tasks, there are still some limitations that need to be addressed, such as the lack of consideration of rich spatio-temporal information in videos and the simplicity of text descriptions used to describe video categories, which results in insufficient contextual description ability. In this paper, we propose a language-video contrastive learning model based on spatio-temporal auxiliary information supervision. For video encoder, we propose a category token-based temporal weighted displacement module for temporal modeling, which enables temporal information to be propagated at various levels of the network from the bottom to the top. Furthermore, we propose a spatio-temporal information auxiliary supervision module to deeply explore the rich spatio-temporal information embedded in visual tokens. For language encoder, we propose a prompt learning method based on large-scale language pre-training models to extend action category text descriptions and generate text descriptions with rich contextual semantic information. The experiment has achieved better results than the current most advanced methods on four video action recognition datasets, namely, mini-Kinetics-200, Kinetics-400, UCF101, and HMDB51, and it is better than or comparable to the current state-of-the-art method, and the accuracy is 2.5%, 0.3%, 0.6% and 2.4% higher than the baseline, respectively.

Keywords action recognition; multimodal model; temporal modeling; spatio-temporal information auxiliary supervision; prompt learning

1 引 言

视频行为识别是计算机视觉领域的热点研究课题之一,在近几十年的时间里引起了众多研究者的关注. 视频行为识别的基本方法被广泛应用于互联网视频审核、视频监控、人机交互等领域.

经典的视觉分类框架,在仅有视觉这一模态的情况下,可以看作是多类别分类任务. 一般地,对于给定的输入视频x及其类别标签y,y是来自于预定义的标签集合Y,模型的参数定义为μ. 经典框架下训练模型时通常是预测条件概率$P(y|x,\mu)$,并将类别标签y转化为数字或者独热向量. 在测试阶段,条件概率的大小代表着置信度分数,最高置信度分数所对应的索引认为是相应的类. 这种框架忽略了标签所描述视觉内容的语义信息,这与人类在理解视频内容时同时利用视觉信息及其对应的语言描述是不同的. 同时,这种训练方式可以看作是一种基于闭集设置的学习方式,在实际问题中对视觉进行自动分类时,有很多新的视觉类别是在模型训练时见不到的,此时使用预定义类别独热向量作为监督信息进行优化的网络模型将无法识别未见过的类别或者不熟悉的类别,极大地限制了模型的泛化能力及其实用价值.

为了解决以上经典分类任务框架中的问题,近年来,研究者们考虑使用类别标签的语言描述作为监督信息来学习新的视觉表达,提出了基于大规模语言-图像样本对训练的多模态模型,以适用于零样本、小样本以及全监督等多种场景. 例如,CLIP模型[1] (Contrastive Language-Image Pre-training),

ALIGN模型[2]（A Large-scale Image and Noisy-text embedding）和Florence模型[3]，这些模型增加了对图像类别语言描述语义信息的利用，使得经典的单模态图像分类框架拓展为基于语言和图像两种模态．多模态模型一般由语言编码和图像编码两个网络模型组成，分别处理语言和图像数据．其中，语言编码网络将图像类别的语言表述通过Transformer进行编码形成语言表达，图像编码将RGB图像通过卷积网络或者视觉Transformer（Vision Transformer，ViT）网络得到视觉表达．编码后的语言表达和视觉表达线性投影到多模态嵌入空间上，接着计算两个模态之间的相似性，使得一个批次内匹配的图文对相似度最大，不匹配的图文对相似度最小．训练时使用对称的交叉熵损失函数，推断时则类似图像匹配任务，以匹配分数最高的类别作为识别结果．

最近的研究工作中，研究者们使用CLIP模型迁移到不同的下游计算视觉任务，例如点云理解[4]，密集预测[5-6]，视频检索[7]等，在各领域达到了当前最优的识别准确率．目前，将CLIP模型应用到视频行为领域还处于初步发展阶段，主要需要解决的问题有以下两个：首先，预训练的语言-图像模型是利用单幅图像而不是视频训练得到的，缺乏时间-空间联合建模能力；其次，描述视频行为类别标签的语言描述过于单调，缺少完整和细致的语言描述刻画上下文语义信息．

为解决上述问题，本文提出了基于时空信息辅助监督的语言-视频对比学习模型，该模型由语言编码网络、视频编码网络和时空信息辅助监督模块构成，网络总体框架图如图1所示．其中，语言编码网络的输入为语言信息经过提示学习和字符对编码（Byte Pair Encoder，BPE）[8]后的等长向量．编码过程为：使用大规模语言预训练模型对行为类别语言描述进行提示学习，得到描述性语句，再经过BPE算法得到等长的编码向量．语言编码网络采用Transformer模型，输出为类别标签的语言表达；视频编码网络则由时序加权位移模块（Temporal Weighted Shift，TWS）、帧内空间注意力（Intra-Frame Spatial Attention，IFSA）模块和前馈网络（Forward Feedback Network，FFN）堆叠多层组成．输入为视频的采样帧图像经过块划分、块嵌入和位置嵌入等操作后形成的序列，输出则为经过时序建模后的视频表达．最后，在语言编码网络和视频编码网络的末端计算这两个表达之间的相似性．对于时空辅助监督模块，则使用视频编码网络输出的视觉词元和类别词元作为输入，视觉词元经过时空聚合得到的表达与类别词元形成的表达进行融合得到视频时空表达，利用该表达执行经典的多类别分类

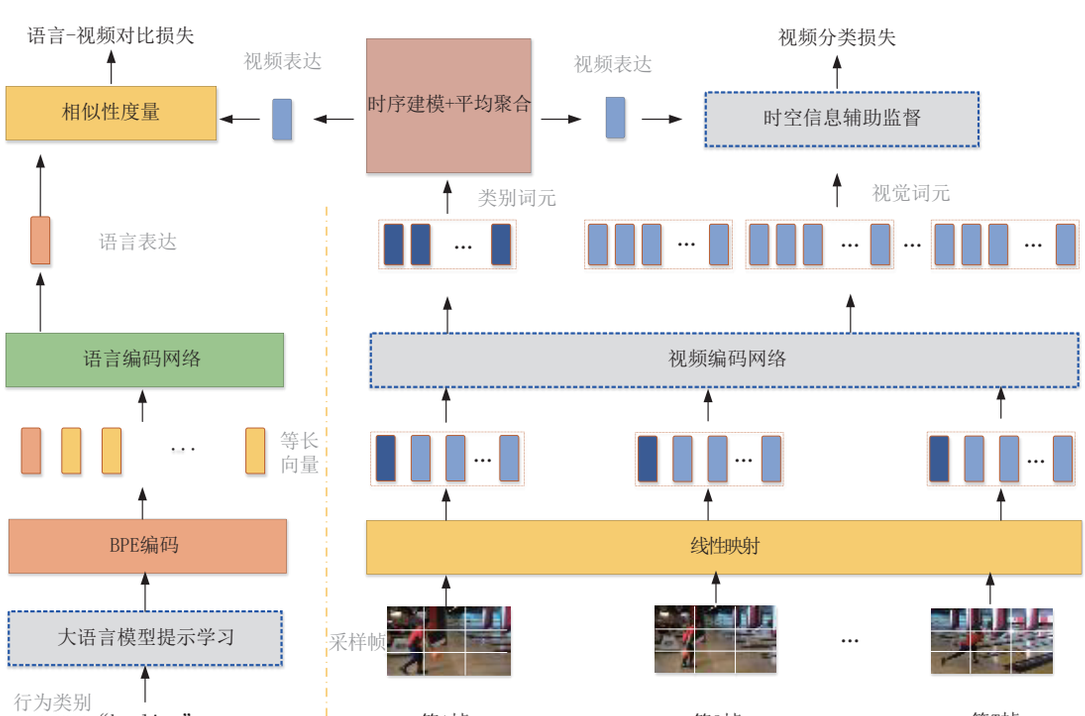

图1　基于时空信息辅助监督的语言-视频对比学习模型网络总体框架图

任务，该模块的作用是辅助监督语言-视频对比学习模型的优化过程.

2 相关工作

2.1 语言-视觉预训练模型

计算机视觉任务在单模态领域取得了较大的进展，但现实世界的问题往往涉及多种模态. 例如，自动驾驶汽车应该能够处理人类指令（语言）、交通信号（视觉）和道路状况（视觉和声音）. 在多模态领域中，视觉与语言的融合备受关注，因为视觉是人类理解环境的最重要感知之一，而与语言相匹配的视觉特征可以极大地提高视觉任务和视觉-语言任务的性能. 在过去的几年时间里，研究者们对于语言-视觉预训练的研究进展迅速，其中具有代表性的是VisualBERT[9]和VideoBERT[10]，分别是第一个语言-图像预训练模型和第一个语言-视频预训练模型. 对比语言-图像预训练模型显示出强大的“零样本”迁移能力和泛化能力，其中CLIP模型[1]是目前最被广泛应用的，CLIP模型利用4亿对语言-图像文本对成功训练了多模态模型，该模型可以应用到多个计算机视觉的下游任务中，例如CoOp[11]，CLIP-Adapter[12]和Tip-Adapter[13]使用预训练的CLIP模型来进行小样本识别任务的迁移，而PointCLIP[4]和DenseCLIP[5]则是利用CLIP模型的知识进行点云理解和密集预测. 受限于视频数据集的规模、质量和计算资源，目前还没有性能较强的语言-视频对比学习预训练大规模网络模型. VideoCLIP[14]在由语言-视频组成的HowTo100M数据集[15]上进行了初步的尝试，然而，这样的视频-文本预训练模型计算成本高，并且需要大量精心标注的语言-视频样本对，不容易采集数据. 所以本文方法直接利用预训练CLIP模型进行拓展用于视频行为识别，节省了训练成本.

2.2 视频行为识别模型

目前视频行为识别领域的基础模型可以分为四类：基于双流卷积网络的行为识别方法[16-20]、基于3D卷积网络的行为识别方法[21-22]、基于2D卷积网络时空建模的行为识别方法[23-26]和基于ViT模型[27]的行为识别方法[28-32].

在双流卷积网络模型中，将视频的表观（RGB）和运动（光流）信息分别作为两个独立网络的输入，并在网络中间层或末端融合双流，是基于深层网络的方法取代以iDT[33]为代表的手工特征方法的里程碑式的网络模型. 3D网络模型通过拓展2D卷积到3D卷积，直接从视频中学习时空联合特征，由于3D卷积网络模型计算代价高，许多计算效率高的网络旨在找到精度和速度之间的平衡. 基于2D卷积网络和时空建模的行为识别方法，其目的在于解决视频网络训练的速度问题，其核心思路是仅采用RGB图像为输入，以2D卷积网络为基础框架，设计有效且计算高效的时空建模模块.

在ViT模型提出之前，基于2D卷积网络和时空建模的方法较为普遍. 其中，TSM[23-24]沿着时间维度移动部分通道，使得视频帧之间的时序信息能够进行交互，从而达到3D卷积的时空建模效果，同时只有2D卷积的复杂度. TEA[34]提出了运动激励模块和多尺度时间聚合模块. STM[35]设计了通道级别的时空模块和动态模块. TDN[36]提出了短时和长时帧间运动变化模块用于获取局部运动信息和全局的运动信息，在速度和精度上都取得了较好的结果，TAM[37]将时序卷积核参数分解成位置敏感的自适应权重和位置无关的自适应卷积核，以自适应方式动态地学习视频中的时序信息. GST[38]同时使用2D和3D卷积学习时空特征，GSM[39]提出了一个空间3D时空分解中的门策略卷积作为特征提取. TEINet[40]包含了一个运动增强模型，用于增强运动相关的特征，同时抑制不相关的信息（如背景）. 同时引入了一个时序交互模块，该模块以信道方式补充时态上下文信息. 以上时空建模方法不仅能够灵活有效地捕捉时间结构，而且能够有效地进行模型推理. 上述工作的共同特点是通过在单个网络中设计基于2D或3D卷积的时空模块，学习帧间的多尺度时序推理关系，建模视频时空信息.

ViT的出现动摇了卷积网络在计算机视觉各个任务上的主导地位. 在视频行为识别领域，基于ViT模型的行为识别方法，基本思路是将ViT模型中仅能建模空间信息的自注意力机制拓展为时空自注意力机制，用于建模视频的时空信息. ViViT[31]、TimeSformer[32]和VidTr[30]是该类方法的代表性工作. 然而，这些方法都是基于RGB图像的单模态方法，忽略了类别标签所具有的语义信息. 基于ViT网络，也有一些基于语言-视觉多模态模型的研究工作，其中主要有9个工作与本文提出的方法最为相关：ActionCLIP[41]、X-CLIP[42]、A6[43]、AIM[44]、ViFi-CLIP[45]、M^2-CLIP[46]、BiKE[47]、DiST[48]和ILA[49]. 其中，ActionCLIP提出了基于预训练CLIP模型[1]的行为识别新框架，对于语言编码网络，其采用手工

模板的方式增强行为类别的表达能力，然而这种方式不能够自适应地获取行为类别语言描述丰富的上下文信息. A6针对语言编码网络设计了增强模块，使得模型可以自适应地学习到行为类别语言描述对应的语义信息. 对于语言编码网络，X-CLIP则提出了利用视频表达和语言表达进行注意力机制的交互来增强行为类别的语义信息. 对于多模态模型的视觉编码网络的时空特征建模问题，ActionCLIP探究了网络中间层输出的视觉词元和类别词元进行基于自注意力机制的时空建模和时序位移，导致了对预训练模型参数的灾难性遗忘问题. X-CLIP则对网络中间层类别词元进行线性变换并对其进行基于自注意力机制时序建模，解决了ActionCLIP在中间层进行时空建模失败的问题. 然而，基于自注意力机制的时序建模计算代价较高，降低了模型效率. AIM和ViFi-CLIP提出了预训练的CLIP用于视频行为识别任务时的策略，例如冻结语言编码网络，学习率的调整等. 上述两种方法可以通过参数初始化的方式利用语言-视觉预训练模型的优势，通过在图像主干下简单地重用预训练的自注意机制进行时间建模，没有充分考虑在视觉分支中对视频中丰富时空信息进行时空建模，也没有考虑在语言分支进行提示学习以丰富类别标签语言描述的上下文信息. M^2-CLIP提出了一种适配模块增强全局信息和建模局部差分运动信息；DiST提出了使用基于轻量级的时序建模模块的时空适配方法；ILA提出了一种基于时序对齐的时空适配模块. BIKE则是在语言分支利用视频属性关联机制引入辅助属性进行提示适配，在视觉分支通过引入基于时间显著性的适配模块增强视频表征的鲁棒性. 上述方法虽然考虑了视觉网络中间层的时空建模，但是忽略了末端视觉词元丰富时空信息的利用.

针对视频数据的特点，为了更好地利用CLIP模型的先验知识，本文在这些工作的基础上提出了基于时空信息辅助监督的语言-视频对比学习模型. 本文提出的模型同时考虑视觉词元的时空信息、视觉编码网络轻量化时空建模和利用大语言模型生成类别标签语言描述的上下文信息，获得了更强的语言监督信息，从而弥补了以上工作的不足. 对于语言编码网络，本文提出了使用大语言模型自适应地获取行为类别语言描述丰富的上下文信息. 同时，对于视频编码网络，受到TSM[23-24]的启发，本文提出了适用于多模态模型的TWS时空建模方法. 本文提出的TWS时序建模方法和TSM方法的主要区别有以下两点：第一，不同于2D CNN网络各阶段的卷积特征仅代表空间局部信息，多模态模型中基于Transformer的视觉编码网络中各阶段特征由类别词元和视觉两类词元构成，分别代表网络某阶段特征的空间全局信息和局部信息. TSM方法对各阶段整体卷积特征进行部分通道时序位移，如果TSM操作直接用于视觉Transformer网络的各个阶段输出的类别词元和视觉词元，则会导致模型性能显著下降，这是由于模型参数的灾难性遗忘导致的[41]. 本文提出的TWS时序建模方法仅对类别词元经过线性变化后形成的通信词元进行部分通道时序位移，相当于只对各阶段特征的空间全局信息进行时序建模，也达到了较好的效果.

对于视觉编码网络末端的时空特征建模，ActionCLIP和X-CLIP都只保留了网络输出的类别词元，并对其进行时序多头自注意力操作后聚合为视频全局表达，忽略了网络输出的视觉词元所包含的丰富的视频局部空间信息. 这些工作初步探究了将基于多模态的对比语言-图像学习的预训练模型应用到视频行为识别任务.

3　时空信息辅助监督的语言-视频对比学习模型

3.1　模型概述

本文提出的基于时空信息辅助监督的语言-视频对比学习模型是CLIP模型用于视频行为识别领域的针对性拓展. 对于给定的一个包含某类行为的视频序列 $V\in V_{\text{all}}$ 及其行为类别语言描述 $I\in I_{\text{all}}$，这里 V_{all} 表示视频的集合，I_{all} 表示目标数据集中所有行为类别语言描述的集合. 将从视频序列 V 中采样的视频帧经过块划分和位置嵌入后送入视频编码网络 g_{θ_v}，得到视频表达 $\boldsymbol{v}$；行为类别语言描述 I 经过大语言模型提示学习后得到 $\hat{I}$，经过编码后进入语言编码网络 g_{θ_c}，得到语言表达 $\boldsymbol{c}$：

$$\boldsymbol{v}=g_{\theta_v}(V),\boldsymbol{c}=g_{\theta_c}(\hat{I}) \tag{1}$$

3.2　视频编码网络时空建模

下面介绍视频编码网络 g_{θ_v}. 具体地，对于一段输入视频序列 V，从中采样 T 帧图像，则 g_{θ_v} 的输入为 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}\in R^{H\times W\times C\times T}$ 的张量，这里 H 和 W 代表视频帧的高和宽，C 为通道数目. 首先，对于视频采样帧中的每张图像 $\boldsymbol{x}_t,t=1,\cdots,T$ 均划分为 N 个分辨率大小为 $P\times P$ 的图像块 $\{\boldsymbol{x}_{t,i}\}_{i=1}^{N}\in R^{P^2C}$，这里 $N=HW/P^2$ 代表划分图像块的数目. 这些小块

$\{\boldsymbol{x}_{t,i}\}_{i=1}^{N}$通过$\boldsymbol{W}\in R^{P^2C\times D}$进行线性映射后，投影为固定长度的向量，称之为视觉词元，视觉词元级联一个可学习的向量即$\boldsymbol{x}_{class}$，称之为类别词元. 则视频编码网络的输入的第t帧为：

$$\boldsymbol{y}_t^{(0)}=[\boldsymbol{x}_{class},\boldsymbol{x}_{t,1}\boldsymbol{W},\boldsymbol{x}_{t,2}\boldsymbol{W},\cdots,\boldsymbol{x}_{t,N}\boldsymbol{W}]+\boldsymbol{E}_{pos},\quad \boldsymbol{W}\in R^{(P^2C)\times D},\boldsymbol{E}_{pos}\in R^{(N+1)\times D} \tag{2}$$

这里$\boldsymbol{E}_{pos}$表示空间位置编码，D为经过线性映射后视频编码网络输入的维度. 经过上述的块嵌入和位置嵌入，视频的采样帧都生成了等长向量，这些向量将作为视频编码网络的输入.

如图2所示，语言-视频对比学习模型中的视频编码网络是由时序加权位移模块(TWS)、帧内空间注意力(IFSA)和前馈网络(FFN)堆叠L层构成的. 首先介绍TWS模块，受到X-CLIP[42]工作的启发，本文方法也利用每一帧图像的类别词元建模视频的时空信息. 第t帧在第l层($l=1,2,\cdots,L$)的类别词元为$\boldsymbol{y}_{t,0}^{(l-1)}$，对其经过线性变换生成新的词元，称之为通信词元：

$$\boldsymbol{m}_t^{(l)}=linear(\boldsymbol{y}_{t,0}^{(l-1)}) \tag{3}$$

第t帧在第l层的通信词元$\boldsymbol{m}_t^{(l)}$能够代表当前帧的全局视觉信息，在视频编码网络中的每一层仅对通信词元进行时序建模是有效的且计算代价较小. 接下来，在第l层利用TWS模块对一段视频的T个采样帧的通信词元进行时序建模，这一过程可以表示为：

$$\hat{\boldsymbol{M}}^{(l)}=\boldsymbol{M}^{(l)}+TWS(\boldsymbol{M}^{(l)}) \tag{4}$$

上式中$\hat{\boldsymbol{M}}^{(l)}=[\hat{\boldsymbol{m}}_1^{(l)},\hat{\boldsymbol{m}}_2^{(l)},\cdots,\hat{\boldsymbol{m}}_T^{(l)}]$表示经过时序建模后的$T$个采样帧图像在第$l$层的通信词元. 如图2所示，TWS模块由两个操作组成：通道卷积和时序位移. 第l层每一帧视频特征的通信词元$\boldsymbol{m}_1^{(l)},\boldsymbol{m}_2^{(l)},\cdots,\boldsymbol{m}_T^{(l)}$都分别经过通道卷积：

$$\tilde{\boldsymbol{M}}^{(l)}=f_\omega(\boldsymbol{M}^{(l)}) \tag{5}$$

在上式中，f_ω是带有参数ω的通道卷积操作，具体实现时，f_ω可采用卷积核大小为1分组数等通道数的1的分组卷积实现. 经过公式(5)的操作，可以获得第l层的通信词元各个通道的权重，得到经过通道加权后的通信词元$\tilde{\boldsymbol{M}}^{(l)}$. 接着，使用无参的时序位移方式对通信词元进行时序建模，$\tilde{\boldsymbol{M}}^{(l)}$进入时序位移模块：

$$\hat{\boldsymbol{M}}^{(l)}=TS(\tilde{\boldsymbol{M}}^{(l)}) \tag{6}$$

在这里，TS表示无参的时序位移操作. 其具体操作过程如图2所示.

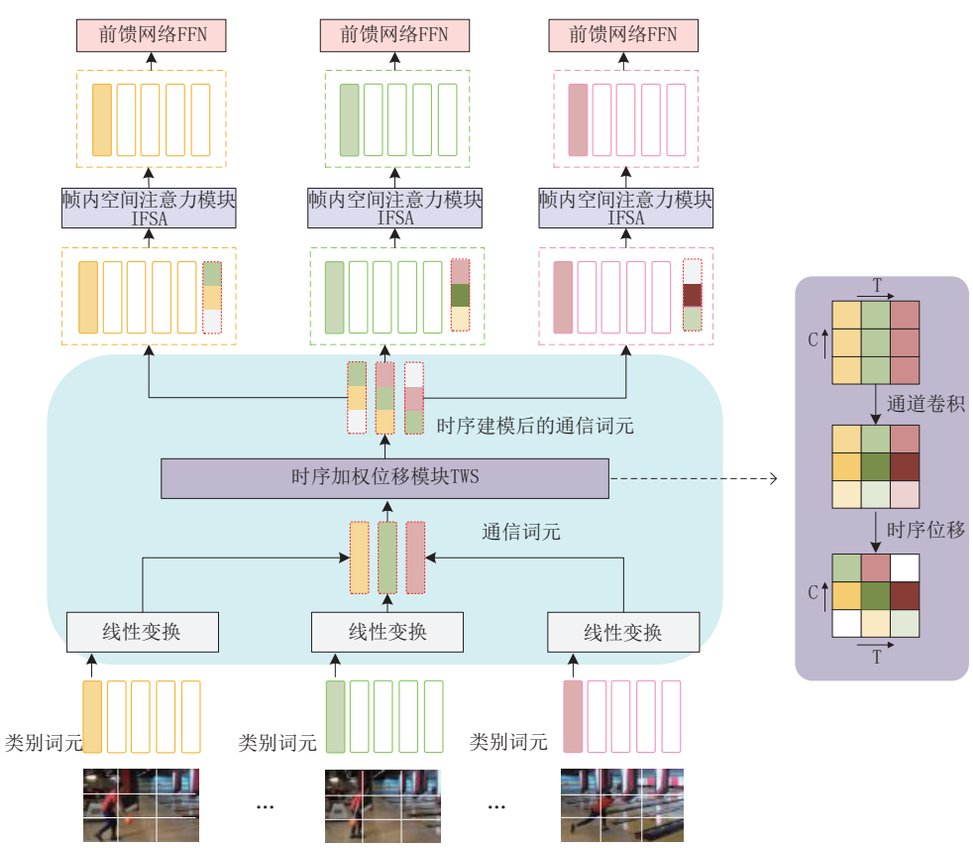

图2 视频编码网络示意图

对于加权后的$\tilde{\boldsymbol{M}}^{(l)}\in R^{T\times D}$，$TS$进行如下操作：首先，将$\tilde{\boldsymbol{M}}^{(l)}$沿通道维度划分为3组：

$$\begin{cases}\tilde{\boldsymbol{M}}^{(l)}=[\boldsymbol{S}_a,\boldsymbol{S}_b,\boldsymbol{S}_c]\\ \boldsymbol{S}_a,\boldsymbol{S}_b,\boldsymbol{S}_c\in R^{T\times a},R^{T\times b},R^{T\times c}\\ a+b+c=D\end{cases}\tag{7}$$

式中$\boldsymbol{S}_a$、$\boldsymbol{S}_b$和$\boldsymbol{S}_c$代表$\tilde{\boldsymbol{M}}^{(l)}$分组后的向量，$a$、$b$和$c$表示分组后向量的维度，$T$表示采样帧数目. 接下来，对于$\boldsymbol{S}_a$和$\boldsymbol{S}_c$组内的通道沿着时序向前和向后移动，对于$\boldsymbol{S}_b$组内的通道元素则保持不变：

$$\begin{cases}\boldsymbol{S}_a(t)=\boldsymbol{S}_a(t-1)\\ \boldsymbol{S}_b(t)=\boldsymbol{S}_b(t)\\ \boldsymbol{S}_c(t)=\boldsymbol{S}_c(t+1)\\ t=1,2,\cdots,T\end{cases}\tag{8}$$

通道移动的比例由a/D和c/D决定，在本文中，$a/D=1/8, c/D=1/8$. 综上所述，TWS模块以最小的计算代价考虑了视频各层T帧通信词元的通道权重和时序信息. 接下来，带有通信词元的特征，进入到IFSA模块中进一步学习视频时空特征：

$$[\hat{\boldsymbol{y}}_t^{(l)},\boldsymbol{m}_t^{(l)}]=[\boldsymbol{y}_t^{(l-1)},\hat{\boldsymbol{m}}_t^{(l)}]+IFSA(\mathrm{LN}([\boldsymbol{y}_t^{(l-1)},\hat{\boldsymbol{m}}_t^{(l)}])\tag{9}$$

上式中，$\boldsymbol{y}_t^{(l-1)}$表示第t帧第$l-1$层输出的类别词元和视觉词元，$[\cdot]$代表将$\boldsymbol{y}_t^{(l-1)}$和第l层通信词元$\hat{\boldsymbol{m}}_t^{(l)}$排在一起. LN代表层归一化操作，IFSA的操作与CLIP模型中视觉Transformer网络的多头自注意力机制(MSA)相同的操作，对于每一帧图像特征对应的词元和通信词元排在一起后形成向量是独立处理的，即在每一帧内部进行注意力机制的计算，帧与帧之间是不进行交互的，所以称之为帧内空间注意力. 最后，将每一帧图像特征对应的词元通过由两层全连接层和GeLU激活函数构成的前馈网络中(FFN)：

$$\boldsymbol{y}_t^{(l)}=\hat{\boldsymbol{y}}_t^{(l)}+FFN(\mathrm{LN}(\hat{\boldsymbol{y}}_t^{(l)}))\tag{10}$$

在上式中可以发现，通信词元并没有进入到FFN中，其作用是仅在当前层进行信息的传递，而不传递到下一层中，这是由于通信词元是在线生成的，可学习的，只影响当前层的信息交互。在TWS模块和IFSA组成的堆叠网络末端，输出为视频每一帧的视觉词元和类别词元，将视频每一帧的类别词元作为最终的表达：

$$\boldsymbol{z}_t=\boldsymbol{y}_{t,0}^{(L)}\tag{11}$$

T帧图像可以得到视频表达$\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_T]$，视频表达则是对这$T$个预测结果在时间维度上计算平均值：

$$\boldsymbol{v}=mean(h_\delta(\boldsymbol{Z}+\boldsymbol{E}_{temp}))\tag{12}$$

上式中$mean$表示沿着时间的维度计算视频表达$\boldsymbol{Z}$的均值，$\boldsymbol{E}_{temp}$代表时序位置编码，h_δ表示时序建模模块，可以是时序Transformer模块或TWS模块等时序建模模块.

3.3　语言编码网络提示学习

本小节介绍语言编码网络g_{θ_c}及本文提出的提示学习方法. 行为类别对应的语言描述$\boldsymbol{I}$，首先通过大语言模型对行为类别语言描述进行提示学习：

$$\hat{\boldsymbol{I}}_{\mathrm{all}}=\boldsymbol{P}_\varphi(\boldsymbol{I}_{\mathrm{all}})\tag{13}$$

这里$\boldsymbol{I}_{\mathrm{all}}$表示目标数据集中的所有行为类别语言描述，$\hat{\boldsymbol{I}}_{\mathrm{all}}$表示经过提示学习后的行为类别语言描述，$\boldsymbol{P}_\varphi$代表提示学习的过程.

图3展示的是通过大规模训练语言模型的API接口生成的行为类别语句描述的示例. 大规模训练语言模型采用的是OPENAI在大规模语料库上训练的系列模型，有8个不同的模型，参数从1.25亿到1750亿不等，本方法中选用的是Text-davinci-002网络模型. 对类别标签进行提示学习是通过问答任务进行的，设置的问题模板是："What are useful visual features for distinguishing＋动作类别名称＋in a video?". 如图3所示，以登山运动中"abseiling"这个动作为例，设置的问题为"在视频中识别'沿绳滑下'需要哪些有用的视觉特征?"大语言模型给出的答案为"沿绳滑下是一种下降形式，登山者使用摩擦装置下降绳索. 登山者通常佩戴安全带，绳索固定在登山者上方的锚点上."

图3　行为类别语言描述的提示学习

对于目标数据集中的行为类别语言描述$\boldsymbol{I}_{\mathrm{all}}$使用上述过程离线地生成$\hat{\boldsymbol{I}}_{\mathrm{all}}$，以增强行为类别语言描

述的上下文语义信息. 对于语言编码网络的输入行为类别语言描述I,根据索引在$\hat{I}_{all}$中找到其经过提示学习后的描述,再利用BPE编码将这些描述信息编码为等长的向量$\tilde{I}$,进而输入到语言编码网络模型中:

$$c=g_{\theta_c}(\tilde{I}) \tag{14}$$

语言模型g_{θ_c}与预训练的CLIP模型一致,采用的是与GPT-2网络模型相同的网络架构,该模型是由多头自注意力模块和FFN层堆叠而成的Transformer网络,网络深度为12层,宽度为512,多头注意力模块中头的个数为8. c为语言编码网络的输出,是其末端代表句子结束标识的EOS词元,也是行为类别语言描述I的语言表达.

得到视频表达v和语言表达c之后,在相似性计算模块,通过计算视频和语言这两个表达之间的余弦距离来获得两种模态之间的对称相似性矩阵,即视频表达和语言表达之间的相似性分数:

$$sim(v,c)=\frac{\langle v,c\rangle}{\|v\|\|c\|} \tag{15}$$

这里$\langle\cdot\rangle$代表内积计算,$sim(\cdot)$代表视频表达v和语言表达c之间的相似度. 接着,该相似性分数经过softmax函数归一化:

$$p_i^{v2c}(v)=\frac{\exp(\mathrm{sim}(v,c_i)/\gamma)}{\sum_{j=1}^{M}\exp(\mathrm{sim}(v,c_j)/\gamma)} \tag{16}$$

其中,γ是一个可学习的温度系数,M代表训练时语言-视频样本对的数目. 令$q_i^{v2c}(\mathrm{v})$表示相似性分数的真值,其矩阵中0元素位置代表负样本对,1元素位置代表正样本对. 接着,定义交叉熵损失CE作为语言-视频之间的对比损失以优化网络:

$$L_m=\boldsymbol{w}\cdot CE\left(p_i^{v2c}(v),q_i^{v2c}(v)\right) \tag{17}$$

这里,L_m表示语言-视频对比学习模型的损失函数,$\boldsymbol{w}$是语言-视频对比学习模型的权重,是一个可调节的参数,在网络训练过程中可以采用固定的标量或者设置为可学习的参数.

3.4 视频编码网络时空信息辅助监督

时空信息辅助监督是本文提出的多模态视频行为模型的重要组成部分,该模块的具体计算流程如图4所示. 如3.2节所述,语言-视频对比学习模型的末端输出是类别词元和由T帧图像的视觉词元构成的视频时空特征,而用于最终表达的词元是类别词元,视觉词元是被抛弃掉的,而本小节提出的辅助监督模块中将同时利用类别词元和视觉词元的重要时空信息,并对视觉词元进行基于幂正规化协方差池化的时空聚合.

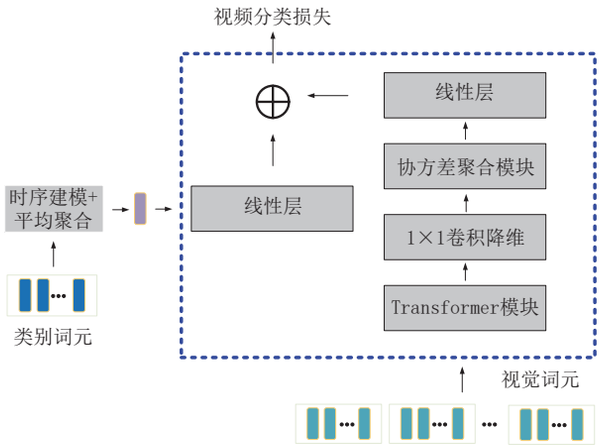

图4 时空信息辅助监督模块示意图

时空信息辅助监督模块的输入为语言-视频对比学习模型的最终视频表达v和每一帧的视觉词元:

$$\boldsymbol{h}_t=[\boldsymbol{y}_{t,1}^{(L)},\boldsymbol{y}_{t,2}^{(L)},\cdots,\boldsymbol{y}_{t,N}^{(L)}] \tag{18}$$

这里,$\boldsymbol{h}_t$表示第t帧图像经过堆叠的TWS模块和IFSA模块之后得到的视觉词元,则$\boldsymbol{H}=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T]\in R^{NT\times D}$表示一段视频中$T$帧图像的视觉词元的集合,这些视觉词元蕴含着每一帧图像中丰富的空间信息. 本小节采用幂正规化协方差池化[50-51]的方法对这些视觉词元进行二阶聚合. 为了与视频编码网络输出的类别词元保持一致,对视觉词元$\boldsymbol{H}$也进行时序Transformer建模,这里的Transformer也是由多头自注意力机制模块和FFN组成,不再赘述,经过Transformer后得到时序建模后的视觉词元$\tilde{\boldsymbol{H}}$,$\tilde{\boldsymbol{H}}$首先经过1×1卷积,对$\tilde{\boldsymbol{H}}$进行降维操作,维度由D维降低到D'维,D'对识别准确率的影响将在实验部分评估. 接下来,协方差池化聚合T帧图像的视觉词元的过程如下:

$$Z_{GCP}=COV\left(\tilde{\boldsymbol{H}}\right) \tag{19}$$

式中COV表示求$\tilde{\boldsymbol{H}}$的协方差,Z_{GCP}则是视频视觉词元的二阶统计信息,对Z_{GCP}进行幂正规化[50-51]操作得到$\hat{Z}_{GCP}$. 同样,本文采用快速迭代法近似协方差矩阵的幂. 最后全局视频表达$\hat{Z}_{GCP}$可以与类别词元的最终表达v融合作为时空信息辅助监督模块的视频全局表达:

$$Z=\mathrm{softmax}(FC(\hat{Z}_{GCP})+FC(v)) \tag{20}$$

式中softmax表示归一化指数函数,FC表示线性层. 假设$\hat{\mathbf{Z}}$为视频对应的标签,则时空信息辅助监督模块的视频分类损失函数为:

$$L_s=CE(Z,\hat{Z}) \tag{21}$$

这里L_s表示时空信息辅助监督模块的损失函数，CE表示交叉熵损失函数. 则本文所提出的方法的损失函数可以表示为：

$$L_{ms}=w\cdot L_m+(1-w)\cdot L_s \tag{22}$$

这里w是语言-视频对比学习模型的损失函数的权重，是一个可调节的参数，能够代表语言-视频对比学习模型的重要性程度，实验部分4.3节的表3(c)将对不同的w取值进行详细的评估. 模型训练时使用对比损失和交叉熵损失，即按照公式(22)实现；测试时则以语言-视觉对比分支为基准，此时将不以辅助监督分支的结果作为最终准确率. 其测试过程类似图像匹配任务，以匹配分数最高的作为识别结果. 具体过程为：编码后的语言表达和视频表达线性投影到多模态嵌入空间上，接着计算两个模态之间的相似性，使得一个批次内匹配的与语言-视频对相似度分数最大，不匹配的图文对相似度分数最小.

4　实　验

4.1　数据集介绍

本节实验将在包括mini-Kintics-200[52]、Kinetics-400[53]、HMDB51[54]和UCF101数据集评估本文提出的方法. 其中，Kinetics-400数据集包含400类行为类别. 其中，训练集样本数目为23 643，验证集样本数目为19 761，视频数据总大小约132 GB. mini-Kintics-200是Kinetics-400数据集的子集，包括200类行为类别，训练集样本数目为77 166，测试集样本数目为4988. UCF101数据集包含101种行为，共13 320个视频序列，这个数据集里面的大多数行为是关于体育运动的. HMDB51数据集包含51种行为，总共6766个视频序列，每一类行为至少涵盖100个视频样本，该数据集视频主要来源于网络视频和电影片段，行为的类内差距非常大，背景复杂，是最具挑战性的数据集之一. 这两个数据集使用3种方式划分训练集和测试集，在3个划分上的平均准确率作为最终的分类结果.

4.2　实验设置

在实验中，本文方法所使用的基础网络架构是在4亿语言-图像对上训练的CLIP模型，其中语言-视频对比学习模型的语言编码网络使用的是深度为12层，宽度为512，有8个注意力头的GPT-2模型. 对于语言-视频对比学习模型的视频编码网络，使用ViT-B/32和ViT-B/16作为基础架构. 对于时空信息辅助监督的语言-视频对比学习模型中不同模块的参数初始化的设置如下：TWS模块中的参数、末端时序建模模块的参数、时空信息辅助监督模块中的参数都是随机初始化的，其他参数都是从CLIP模型初始化的.

在训练本文提出的多模态模型时，网络优化器为带权重衰减的自适应矩估计(AdamW)，批量样本大小设置为256，动量设置为0.9，权重衰减是5e-4. 网络训练分为两个阶段，第一阶段不插入中间层对通信词元时序建模的TWS模块，学习模型中的时空信息辅助监督模块新增加的参数，初始化学习率设置为8e-4，使用余弦学习率，迭代20 epochs. 在第二阶段时加入TWS模块，并更新模型中所有的参数，初始化学习率为8e-6，采用余弦学习率，学习率预热为5 epochs，迭代30 epochs. 对于TWS模块中的参数倍乘1000，除CLIP预训练模型中的参数不进行学习率倍乘外，其他的模型参数都倍乘100. 为了避免过拟合现象的产生，在时空信息辅助监督模块中视觉词元的时空聚合层之后插入dropout层. 在所有消融实验中，采用均匀采样方法，即将视频均匀分成T个部分，从T个部分中随机采样一帧. 图像增强是对输入的图像进行裁剪，裁剪尺寸的高和宽是在{256；224；192；168}这四个数值中随机选择，然后再把裁剪好的区域改变到224×224.

在测试模型时，对于4.3节所有的消融实验，都是使用ViT-B/32作为基础模型，在mini-Kinetics-200数据集上进行的，比较的是在验证集上的Top-1的单视角测试识别准确率. 对于表5、表6和表7中的与前沿的视频行为识别算法比较的实验结果，则是采用与X-CLIP[42]相同的12个视角测试方法，即从视频中采样4组不同视频片段，并将每组视频中的视频帧沿着长边均匀裁成3份，形成12组不同视频，最终预测结果为这12个视角预测的平均.

4.3　消融实验

(1)视频编码网络中通信词元时序建模方式评估. 如3.3节所述，在视频编码网络中，采用时序加权位移(TWS)的方法对网络各层中采样帧特征的通信词元进行时序建模. 为了证明本文提出的视频编码网络中间层时序建模方法TWS的有效性，在本小节中以语言-视频对比学习模型为基础模型，在没有使用时空信息辅助监督模块的情况下，评估几种不同时序建模方式. 这几种时序建模方式可以看作是在CLIP(基线方法2)的基础上加入的模块，分

别为:时序卷积(TC)、时序位移(TS)和Transformer(Transf.). TC、TS和Transf.分别表示对视频编码网络中的通信词元进行时序卷积建模、时序位移和时序Transformer建模.在TC模块中,使用的卷积核大小为3,在TS模块中通道移动比例为1/8,Tranf.模块则与X-CLIP中的实现一致,采用多头注意力机制和FFN组成的Transformer模块,注意力机制的头数目为8.

如表1中的结果所示.基线方法1和基线方法2分别指的是仅使用预训练CLIP模型的视觉编码网络以及使用CLIP预训练模型在mini-Kinetics-200数据集上进行微调的结果,在这两个基线方法中,视觉编码网络输入的视频采样帧独立进入网络,获得每一帧的最终表达,所有视频表达为所有采样帧表达的均值.基线方法1在mini-Kinetics-200数据集上的Top-1准确率是79.69%,基线方法2的结果为83.42%,由此可见,增加了语言监督的多模态模型,相比于仅有视觉模态的网络模型,使得Top-1准确率提升3.73%.这初步验证了多模态模型在视频行为识别任务上的有效性,因此之后的所有实验都在多模态模型的情况下进行,基线方法2是后续对比实验中的基线方法.

表1 视频编码网络通信词元时序建模方式评估

方法	模态	时序建模方法	GFLOPs	Top-1准确率(%)
CLIP-视觉网络(基线方法1)	视觉	×	35.2	79.69
CLIP(基线方法2)	语言+视觉	×	+0.2	83.42
CLIP(基线方法2)	语言+视觉	TC	+0.4	84.10
CLIP(基线方法2)	语言+视觉	TS	+0.2	84.14
CLIP(基线方法2)	语言+视觉	Transf.	+2.2	84.28
本文方法	语言+视觉	**TWS**	**+0.3**	**84.35**

由表1可知,与基线方法2相比,所有方法均有提升.本文提出的在视频编码网络中进行时序建模的TWS方法比基线方法2提升0.93%.作为TWS模块的对比方法,TC、TS和Tranf.的方法分别比视频编码网络不加入任何时序建模网络模块的基线方法2提升0.68%、0.72%和0.86%.表1中的后5行对比了在基线方法1的GFLOPs的基础上,不同模块的GFLOPs增加的情况,基线方法2比基线方法1增加的GFLOPs为0.2,TS模块与其增加的相同,即在基于语言和视觉的多模态情况下不增加GFLOPs,但是TS模块的Top-1识别准确率最低,对于TWS、TC和Tranf.模块,增加的GFLOPs分别是0.3、0.4和2.2,由此可以看到本文提出的TWS模块有着较为明显的优势,证明了TWS的有效性,也证明了TWS模块是性能和计算代价的最佳平衡方案.

(2)视频编码网络末端类别词元时序建模的评估.如表2所示,评估了末端类别词元不进行时序建模和使用Transf.进行时序建模的结果.末端类别词元进行Transf.时序建模时,识别准确率为85.12%,比末端不进行时序建模的结果提升0.77%,同时比基线方法2提升1.70%,说明本文方法在视频编码网络中间层对通信词元进行时序建模与末端对类别词元进行Transf.时序建模的方式互补性良好.

表2 视频编码网络末端类别词元时序建模评估

方法	视频编码网络末端时序建模	Top-1准确率(%)
CLIP(基线方法2)	×	83.42
本文方法	×	84.35
本文方法	**Transf.**	**85.12**

表3时空信息辅助监督模块的关键参数评估.在表2实验的基础上,表3(a)(b)(c)中评估了语言-视频对比学习模型增加了时空信息辅助监督模块后的识别准确率以及其中关键参数对性能的影响.基于时空信息辅助监督模块的输入可以是仅有视觉词元,也可以是类别词元和视觉词元,所以在表3(a)中本小节评估了这两种不同的情况.在表3(a)中,具体的参数设置如下:语言-视频对比学习模型和时空信息辅助监督模块的权重系数均为0.5,二阶表达的维度2K,即视觉词元的维度由512维降到64维之后进行聚合的二阶表达.由表3(a)可知,在不使用时空信息辅助模块时,本文方法的识别准确率为85.12%,增加了时空信息辅助监督模块后,识别准确率提升0.35%,这说明了辅助监督模块的有效性.同时,融合了视觉词元的一阶池化表达和类别词元的方法Top-1准确率为85.19%,融合了视觉词元的二阶表达和类别词元的方法为85.47%,仅使用视觉词元的二阶表达作为视频表达的Top-1准

表3　时空信息辅助监督模块的关键参数评估

表3(a)　时空信息辅助监督模块中类别词元和视觉词元使用方式对比

方法	语言-视频对比学习模型		辅助监督		Top-1准确率(%)
	类别词元	视觉词元	类别词元	视觉词元	
CLIP(基线方法2)	√	×	-	-	83.65
本文方法	√	×	-	-	85.12
本文方法	√	×	×	√	85.38
本文方法	√	×	√	√(一阶池化)	85.19
本文方法	√	×	√	√	**85.47**

表3(b)　时空信息辅助监督模块中表达维度评估

方法	降维维度(D')	表达维度	Top-1准确率(%)
CLIP(基线方法2)	-	512	83.65
本文方法	64	2K	85.47
本文方法	96	4K	85.24
本文方法	**128**	**8K**	**85.61**
本文方法	192	18K	85.57
本文方法	256	32K	85.22

表3(c)　时空信息辅助监督模块中权重系数评估

方法	时空信息辅助监督模块权重($1-w$)	Top-1准确率(%)
CLIP(基线方法2)	-	83.65
本文方法	0.7	85.33
本文方法	0.6	85.32
本文方法	**0.5**	**85.61**
本文方法	0.4	85.55
本文方法	0.3	85.28

确率为85.38%，这说明在辅助监督模块中，最终的视频表达融合类别词元和视觉词元能够提升识别准确率，对视觉词元进行协方差聚合的效果优于一阶池化，所以之后的消融实验均采用这种方式.

表3(b)为本文提出的多模态模型中时空信息辅助监督的视频表达维度评估. 在本文提出的多模态模型中，二阶表达的维度是可能影响模型识别准确率的关键参数，在本小节的实验中评估了二阶视频表达的维度，其他参数设置如下：时空辅助监督模块的权重为0.5. 表3(b)中列出了时空信息辅助监督模块中视觉词元经过时空聚合后的二阶表达维度为2K、4K、8K、12K、18K和32K时的Top-1识别准确率，其中不同的表达维度对应的特征在进行时空聚合之前，维度分别是从512降到64、96、128、192和256. 由表3(b)可知，当视觉词元经过时空聚合后的二阶表达维度为8K，即特征的维度从512降到128时，Top-1准确率最高为85.61%. 本文后续的实验中，都采用二阶表达为8K，即特征维度从512降到128，此时本文方法比基线方法2提升1.96%. 表3(c)为本文提出的多模态模型中时空信息辅助监督模块权重系数评估. 以下评估语言-视频对比学习模型损失函数的权重，本小节中的其他参数根据前述消融实验结果选择最优设置. 如表3(c)所示，本小节评估了时空信息辅助监督模块的权重为0.7、0.6、0.5、0.4、0.3时的识别准确率. 当时空信息辅助监督模块的权重系数为0.5时，本文方法识别准确率达到最优，本文方法的Top-1准确率分别比基线方法2提升1.96%. 这说明语言-视频对比学习模型与时空信息辅助监督模块的重要性相同.

(4)不同提示学习方法对比. 对于多模态模型中的语言编码网络的研究，现有的工作中大多数都是对类别标签语言描述进行提示学习方面的改进. 在本小节中，将本文3.3节中的提出的提示学习方法与3种不同的提示学习方法进行了对比. 如表4所示，表中的比较分为两部分，第一部分是在基线方法的情况下，对比是否在语言编码网络中使用本文提示学习时的性能，本文提示学习方法使得基线方法提升0.37%，初步证明了本文对于语言编码网络提出的提示学习方法有效性. 第二部分是本文方法

表4 语言编码网络中不同提示学习方法对比

方法	提示学习方法	Top-1准确率(%)
CLIP(基线方法2)	无	83.65
CLIP(基线方法2)	**本文提示学习方法**	**84.02**
本文方法	无	85.61
本文方法	X-CLIP[42]	85.71
本文方法	CoOp[55]	85.53
本文方法	ActionCLIP[41]	85.49
本文方法	**本文提示学习方法**	**85.93**

采用不同提示学习方法的对比.本文方法在语言编码网络中不采用任何提示学习方法时,Top-1准确率是85.61%.本文方法与X-CLIP[42]中的Video-specific提示学习方法、CoOp[55]以及ActionCLIP[41]中的手工模板方法结合时,Top-1准确率分别为85.71%、85.53%和85.49%,提升幅度较小或无明显提升.当本文提示学习方法采用3.3节所述提示学习方法时,与不使用提示学习方法的本文方法的Top-1准确率相比提升0.32%,比基线方法的Top-1准确率提升2.28%,说明使用大语言模型对行为类别语言描述进行提示学习的方法能够进一步提升本文方法的识别准确率.

4.4 本文方法与其他行为识别方法比较

为了进一步证明本文方法的有效性,本节中将本文方法在mini-Kinetics-200、Kinetics-400、UCF101和HMDB51这4个数据集上与当前视频行为识别中前沿的方法进行对比.表5是在mini-Kinetics-200数据集上本文方法与当前其他方法比较.

表5 Mini-Kinetics-200数据集本文方法与其他方法比较

方法	基础模型	帧数	Top-1准确率(%)
MARS[56]	3D RX101	8	72.8
BAT[57]	2D Slow50	8	70.6
RMS[58]	3D R101	8	78.6
CGNL[59]	3D R101	8	79.5
Ada3D[60]	MobileNet V2	16	79.2
TCPNet[61]	2D R50	8	80.7
V4D[62]	3D R50	4	80.4
DSANet[63]	3D R50	16	81.8
本文方法	**ViT-B/32**	**8**	**86.5**
本文方法	**ViT-B/16**	**8**	**89.6**

如表5所示,表中比较的方法基础模型包括2D和3D卷积网络模型,比较的输入帧数包括4帧,8帧和16帧.在所有基于8帧输入的比较中,基于ViT-B/32和基于ViT-B/16的本文方法比其中最具竞争力的方法TCPNet识别准确率分别高5.8%和8.9%.基于ViT-B/32和基于ViT-B/16的本文方法在仅采用8帧的情况下,分别比最具竞争力16帧方法DSANet的Top-1准确率提升4.7%和7.8%.表5中的比较方法均为基于单模态的方法,这说明了多模态模型的优势十分明显.

表6是在Kinetics-400数据集上本文方法与当前其他行为识别方法比较,展示了目前行为识别研究中基于4种不同网络架构的有竞争力的行为识别方法结果,对比的参数有:预训练数据集、帧数、Top-1准确率、视角、GFLOPs.表6的第一部分是基于3D卷积网络的方法,第二部分是基于2D卷积网

表6 Kinetics-400数据集上本文方法与其他视频行为识别方法比较

方法	预训练	帧数	Top-1准确率(%)	视角	GFLOPs
SlowFast[64]	IN-1k	16+64	74.7	10×3	284
X3D-XXL[65]	IN-1k	16	80.4	10×3	48.4
TPN[22]	IN-1k	32	78.9	10×3	-
CorrNet[66]	IN-1k	32	79.2	10×3	224
SmallBigNet[25]	IN-1k	32	77.4	10×3	418
TEA[34]	IN-1k	16	76.1	10×3	70
TANet[37]	IN-1k	16	79.3	10×3	86
TEINet[40]	IN-1k	16	76.2	10×3	66
VTN-ViT-B[29]	IN-21k	250	79.8	10×3	4218
STAM[31]	IN-21k	64	80.5	10×3	270
MViT-B[67]	-	64	81.2	3×3	455
Uniformer-B[68]	IN-1k	32	83.0	4×3	259
TimeSformer-L[32]	IN-21k	96	80.7	1×3	2380
Mformer-HR[69]	IN-21k	16	81.1	10×3	959
Swin-L[70]	IN-21k	32	83.1	4×3	604
ActionCLIP-B/16[41]	CLIP-400M	32	83.8	10×3	563

（续表）

方法	预训练	帧数	Top-1准确率(%)	视角	GFLOPs
A6[43]	CLIP-400M	8	76.9	-	141
X-CLIP-B/32[42]	CLIP-400M	8	80.4	4×3	39
X-CLIP-B/32[42]	CLIP-400M	16	81.1	4×3	75
X-CLIP-B/16[42]	CLIP-400M	8	83.8	4×3	145
X-CLIP-B/16[42]	CLIP-400M	16	84.7	4×3	287
AIM-ViT-B/16[44]	CLIP-400M	16	84.5	1×3	202
ViFi-CLIP-B/16[45]	CLIP-400M	16	83.9	4×3	141
M^2-CLIP-B/16[46]	CLIP-400M	8	83.4	4×3	214
M^2-CLIP-B/16[46]	CLIP-400M	16	83.7	4×3	422
ILA-ViT-B/16[49]	CLIP-400M	8	84.0	4×3	149
ILA-ViT-B/16[49]	CLIP-400M	16	85.7	4×3	295
本文方法-B/32	**CLIP-400M**	**8**	**80.6**	4×3	**38**
本文方法-B/32	**CLIP-400M**	**16**	**81.3**	4×3	**74**
本文方法-B/16	**CLIP-400M**	**8**	**84.1**	4×3	**142**
本文方法-B/16	**CLIP-400M**	**16**	**84.9**	4×3	**285**

络和时空建模的方法，第三部分是基于Transformer的方法，第四部分是基于多模态模型的方法. 这些模型的预训练数据集有ImageNet-1K（IN-1k）、ImageNet-1K（IN-21k）、JFT-300M、JFT-3B、CLIP-400M、WTS和FLD-900M. 比较的是验证集上的Top-1识别准确率. 如表6所示，基于ViT-B/32和ViT-B/16模型，输入帧数为8的本文方法在识别准确率和计算代价上优势十分明显的. 基于ViT-B/16模型的本文方法识别准确率比基础模型采用ViT-H模型的Swin-L高1.0%. 对于多模态模型的方法，基于ViT-B/16模型，仅采用8帧的本文方法比32帧的ActionCLIP识别准确率高0.3%；基于ViT-B/32和ViT-B/16模型的8帧的本文方法分别比XCLIP识别准确率高0.2%和0.3%. 基于ViT-B/16模型的16帧的本文方法比ViFi-CLIP识别准确率高0.2%. 输入帧数量为16的ViT-B/16模型的本文方法比AIM方法识别准确率高0.4%. 基于ViT-B/16模型的8帧和16帧的本文方法分别比M^2-CLIP-B/16识别准确率高0.7%和1.2%. 基于ViT-B/16模型的16帧的本文方法虽然比ILA-ViT-B/16识别准确率低0.8%，但是ILA-ViT-B/16的GFLOPs比本文方法高10，计算代价较高. 以上结果表明，本文方法是十分有竞争力.

表7中给出了使用Kinetics-400上预训练的视频编码基于ViT-B/32和ViT-B/16的本文方法进行规模较小数据集UCF101和HMDB51数据集微调的结果，比较的方法覆盖双流卷积网络、3D卷积网络、2D卷积时空建模网络和Transformer这4种不同的视频识别网络类型. 从帧数、基础模型和验证集上的Top-1准确率三个方面对本文方法和其他行为识别方法进行了对比.

由表7可知，仅采用输入帧数为8时，基于ViT-B/32的本文方法，优势就十分明显，除了基于16帧的TEA、基于16帧的STM和基于32帧的ViT-L的ViDTr外，本文方法在UCF101数据集上的结果为95.7%，均高于其他方法. 当采用输入帧数为16时，本文方法性能达到97.1%，均高于其他方法. 在

表7　UCF101和HMDB51数据集上本文方法与其他视频行为识别方法比较

方法	基础模型	帧数	UCF101(%)	HMDB51(%)
TSM[23, 24]	2D R50	8	95.2	72.0
ECO[71]	Inc.+3D R18	16	92.8	68.5
TEA[34]	2D R50	16	96.9	73.3
STM[35]	2D R50	16	96.2	72.2
STC[72]	3D RX101	16	92.3	65.4
ART[73]	3D R18	16	94.3	70.9
I3D[21]	3D Inc.	64	95.4	**74.5**
ABM[74]	3D Inc.	64	95.1	72.7
R3D[75]	3D R50	16	92.9	69.4
TCP[61]	2D R50	8	95.1	72.5
MCL[76]	R(2+1)D	16	93.4	69.1
ViDTr[30]	ViT-L	32	96.7	74.4
本文方法	**ViT-B/32**	**8**	**95.7**	**73.8**
本文方法	**ViT-B/32**	**16**	**97.1**	**75.3**
本文方法	**ViT-B/16**	**8**	**97.5**	**76.9**

HMDB51数据集上，基于ViT-B/32的本文方法比64帧的I3D低0.7%，一方面是由于I3D帧数较多，另一方面是I3D采用双流3D卷积网络，计算代价较高. 基于ViT-B/32的本文方法比32帧的ViDTr低0.6%，ViDTr的基础模型ViT-L的参数和GFLOPs约为ViT-B/32的16倍. 对于基于ViT-B/16的本文方法，在UCF101数据集和HMDB51数据集上都取得了当前最好的结果，在UCF101数据集上，8帧的方法比当前最好的结果16帧的TEA高0.6%，在HMDB51上，8帧的本文方法比基于双流3D卷积网络的64帧I3D方法高2.4%. 以上结果进一步表明本文方法在小规模数据集的泛化能力以及针对于不同行为识别任务的有效性.

5 结 论

本文主要在基于语言-图像对比学习的预训练CLIP模型的基础上，提出了一种基于时空信息辅助监督的语言-视频对比学习模型. 本文的核心研究目标是如何拓展CLIP模型用于视频行为识别任务. 首先，对于视频编码网络，本文提出了时序加权位移模块，嵌入到视频编码网络中进行层次化的时序建模；同时，为了深入挖掘网络末端视觉词元所携带的空间信息，通过对其进行二阶聚合生成视觉时空表达，并利用该表达执行分类任务，帮助多模态模型的优化. 其次，对于语言编码网络，使用大语言模型生成行为类别语言描述的上下文信息，使得预训练模型更加适应下游的视频识别任务. 最后，本文实验部分进行了详尽的消融以评估各个模块的作用，并在四个视频识别通用数据集上验证本文方法的有效性，实验结果表明，本文提出的基于时空信息辅助监督的语言-视频对比学习模型在当前基于四种不同架构的视频行为识别算法中都是十分具有竞争力的.

致 谢 感谢国家自然科学基金(61971086、61972062)、辽宁省应用基础研究计划项目(2023JH2/101300191、2023JH2/101300193)、吉林省科技厅科技发展计划项目(20230201111GX)的资助. 感谢审稿专家和编辑在百忙之中审阅本文!

参 考 文 献

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning, Virtual Event, 2021: 1-16

[2] Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the International Conference on Machine Learning, Virtual Event, 2021: 4904-4916

[3] Yuan L, Chen D, Chen Y L, et al. Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432, 2021

[4] Zhang R, Guo Z, Zhang W, et al. PointCLIP: Point cloud understanding by CLIP//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 2022: 8542-8552

[5] Rao Y, Zhan W, Chen G, et al. DenseCLIP: Language-guided dense prediction with context-aware prompting//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event: IEEE, 2022: 18061-18070

[6] Zhou C, Loy C C, Dai B. Extract free dense labels from CLIP//Proceedings of the European Conference on Computer Vision, Virtual Event, 2022: 696-712

[7] Fang H, Xiong P, Xu L, et al. CLIP2Video: Mastering video-text retrieval via image CLIP. arXiv preprint arXiv.2106.11097, 2021

[8] Rico S, Barry H, Alexandra B. Neural machine translation of rare words with sub-word units//Proceedings of the Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 1715-1725

[9] Li L H, Yatskar M, Yin D, et al. VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019

[10] Sun C, Myers A, Vondrick C, et al. VideoBERT: A joint model for video and language representation learning//Proceedings of the International Conference on Computer Vision, Seoul, Korea, 2019: 7464-7473

[11] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models. International Journal of Computer Vision, 2022,130(9): 2337-2348

[12] Gao P, Geng S, Zhang R, et al. CLIP-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv: 2110.04544, 2021

[13] Zhang R, Fang R, Zhang W, et al. Tip-Adapter: Training-free CLIP-adapter for better vision-language modeling//Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 2021: 493-510

[14] Xu H, Ghosh G, Huang P Y, et al. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2021: 6787-6800

[15] Miech A, Zhukov D, Alayrac J B, et al. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips//Proceedings of the International Conference on Computer Vision, Seoul, Korea, 2019

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

[17] Karen S，Andrew Z. Two-stream convolutional networks for action recognition in videos//Proceedings of the Advances in Neural Information Processing Systems，Montréal，Canada，2014：568-576

[18] Wang L，Xiong Y，Zhe W，et al. Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，41(11)：2740-2755

[19] Wang L，Xiong Y，Wang Z，et al. Temporal segment networks：Towards good practices for deep action recognition//Proceedings of the European Conference on Computer Vision，Amsterdam，Netherlands，2016：20-36

[20] Zhang Bing-Bing，Li Pei-Hua，Sun Qiu-Le. Spatial and temporal features aggregation convolutional network model based on locality-constrained affine subspace coding. Chinese Journal of Computers，2020，43(9)：1-15 (in Chinese)
(张冰冰，李培华，孙秋乐. 基于局部约束仿射子空间编码的时空特征聚合卷积网络模型. 计算机学报，2020，43(9)：1-15)

[21] Carreira J，Zisserman A. Quo vadis，Action recognition? A new model and the Kinetics dataset//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Honolulu，USA，2017：4724-4733

[22] Yang C，Xu Y，Shi J，et al. Temporal pyramid network for action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Virtual Event，2020：591-600

[23] Lin J，Gan C，Han S. TSM：Temporal shift module for efficient video understanding//Proceedings of the International Conference on Computer Vision，Seoul，Korea，2019：7082-7092

[24] Lin J，Gan C，Wang K，et al. TSM：Temporal shift module for efficient and scalable video understanding on edge devices. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，44(5)

[25] Li X，Wang Y，Zhou Z，et al. SmallBigNet：Integrating core and contextual views for video classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Virtual Event，2020：1089-1098

[26] Pu Zhan-Xing，Ge Yong-Xin. Few-shot action recognition in video based on multi-feature fusion. Chinese Journal of Computers，2023，46(3)：594-608 (in Chinese)
(蒲瞻星，葛永新. 基于多特征融合的小样本视频行为识别算法. 计算机学报，2023，46(3)：594-608)

[27] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16x16 words：Transformers for image recognition at scale//Proceedings of the International Conference on Learning Representations，Virtual Event，2021：1-21

[28] Arnab A，Dehghani M，HheigoldEIGOLD G，et al. ViViT：A video vision transformer//Proceedings of the International Conference on Computer Vision，Virtual Event，2021：6816-6826

[29] Neimark D，Bar O，Zohar M，et al. Video transformer network//Proceedings of the International Conference on Computer Vision Workshops，Virtual Event，2021：3156-3165

[30] Li X，Zhang Y，Liu C，et al. VidTr：Video transformer without convolutions//Proceedings of the International Conference on Computer Vision，Virtual Event，2021：13557-13567

[31] Sharir G，Noy A，Zeinik-manor L. An image is worth 16x16 words，what is a video worth? arXiv preprint arXiv：2103.13915，2021

[32] Bertasius G，Wang H，Torresani L. Is space-time attention all you need for video understanding? //Proceedings of the International Conference on Machine Learning，Virtual Event：ACM，2021：813-824

[33] Wang H，Schmid C. Action recognition with improved trajectories//Proceedings of the International Conference on Computer Vision，Sydney，Australia，2013：3551-3558

[34] Li Y，Ji B，Shi X，et al. TEA：Temporal excitation and aggregation for action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Virtual Event，2020：906-915

[35] Jiang B，Wang M，Gan W，et al. STM：Spatiotemporal and motion encoding for action recognition//Proceedings of the International Conference on Computer Vision，Seoul，Korea，2019：2000-2009

[36] Wang L，Tong Z，Ji B，et al. TDN：Temporal difference networks for efficient action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Virtual Event，2021：1895-1904

[37] Liu Z，Wang L，Wu W，et al. TAM：Temporal adaptive module for video recognition//Proceedings of the International Conference on Computer Vision，Virtual Event，2021：13688-13698

[38] Luo C，Yuille A. Grouped spatial-temporal aggregation for efficient action recognition//Proceedings of the International Conference on Computer Vision，Seoul，Korea，2019：5511-5520

[39] Sudhakaran S，Escalera S，Lanz O. Gate-shift networks for video action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Seattle，USA，2020：1099-1108

[40] Liu Z，Luo D，Wang Y，et al. TEINet：Towards an efficient architecture for video recognition//Proceedings of the AAAI Conference on Artificial Intelligence，New York，USA，2020：11669-11676

[41] Wang M，Xing J，Liu Y. ActionCLIP：A new paradigm for video action recognition. arXiv preprint arXiv：2109.08472，2021

[42] Ni B，Peng H，Chen M，et al. Expanding language-image pretrained models for general video recognition//Proceedings of the European Conference on Computer Vision，Tel Aviv，Israel，2022：1-18

[43] Ju C，Han T，Zheng K，et al. Prompting visual-language models for efficient video understanding//Proceedings of the European Conference on Computer Vision，Virtual Event，2021：105-124

[44] Yang T，Yi Z，Xie Y，et al. AIM：Adapting image models for efficient video action recognition//Proceedings of the International Conference on Learning Representations，Kigali，Republic of Rwanda，2023：1-18

[45] Rasheed H A K M. Fine-tuned CLIP models are efficient video learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，Vancouver，Canada，2023：6545-6554

[46] Wang M, Xing J, Boyuan J, et al. M2-CLIP: A multimodal, multi-task adapting framework for video action recognition// Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, Canada: AAAI, 2024: 1-9

[47] Wu W, Wang X, Luo H, et al. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 6620-6630

[48] Qing Z, Zhang S, Huang Z, et al. Disentangling spatial and temporal learning for efficient image-to-video transfer learning// Proceedings of the International Conference on Computer Vision, Paris, France, 2023: 13888-13898

[49] Tu S, Dai Q, Wu Z, et al. Implicit temporal modeling with learnable alignment for video recognition//Proceedings of the International Conference on Computer Vision, Paris, France, 2023: 19879-19890

[50] Li P, Xie J, Wang Q, et al. Is second-order information helpful for large-scale visual recognition? //Proceedings of the International Conference on Computer Vision, Venice, Italy, 2017: 2089-2097

[51] Li P, Xie J, Wang Q, et al. Towards faster training of global covariance pooling networks by iterative matrix square root normalization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 947-955

[52] Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification//Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 318-335

[53] Kay W, Carreira J, Simonyan K, et al. The Kinetics human action video dataset. arXiv preprint arXiv: 1705.06950, 2017

[54] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition//Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 2011: 2556-2563

[55] Zhou K, Yang J, Loy C C, et al. Learning to prompt for vision-language models//Proceedings of the International Journal of Computer Vision, Virtual Event, 2022

[56] Crasto N, Weinzaepfel P, ALAHARI K, et al. MARS: Motion-augmented RGB stream for action recognition// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 7874-7883

[57] Chi L, Yuan Z, Mu Y, et al. Non-local neural networks with grouped bilinear attentional transformers//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 2020: 11801-11810

[58] Kim J, Cha S, Wee D, et al. Regularization on spatio-temporally smoothed feature for action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 2020: 12100-12109

[59] Yue K, Sun M, Yuan Y, et al. Compact generalized non-local network//Proceedings of the Advances in Neural Information Processing Systems, Montréal, Canada, 2018: 6511-6520

[60] Li H, Wu Z, Shrivastava A, et al. 2D or not 2D? Adaptive 3D convolution selection for efficient video recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 2021: 6151-6160

[61] Gao Z, Wang Q, Zhang B, et al. Temporal-attentive covariance pooling networks for video recognition//Proceedings of the Advances in Neural Information Processing Systems, Virtual Event, 2021: 13587-13598

[62] Zhang S, Guo S, Huang W, et al. V4D: 4D convolutional neural networks for video-level representation learning// Proceedings of the International Conference on Learning Representations, Virtual Event, 2020: 1-10

[63] Wu W, Zhao Y, Xu Y, et al. DSANet: Dynamic segment aggregation network for video-level representation learning// Proceedings of the ACM International Conference on Multimedia, Virtual Event, 2021: 1903-1911

[64] Feichtenhofer C, Fan H, Malik J, et al. SlowFast networks for video recognition//Proceedings of the International Conference on Computer Vision, Seoul, Korea, 2019: 6201-6210

[65] Feichtenhofer C. X3D: Expanding architectures for efficient video recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 2020: 200-210

[66] Wang H, Du T, Torresani L, et al. Video modeling with correlation networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 352-361

[67] Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers//Proceedings of the International Conference on Computer Vision, Virtual Event, 2021: 6804-6815

[68] Li K, Wang Y, Zhang J, et al. UniFormer: Unifying convolution and self-attention for visual recognition//Proceedings of the International Conference on Learning Representations, Virtual Event, 2022: 1-18

[69] Patrick M, Campbell D, Aano Y M, et al. Keeping your eye on the ball: Trajectory attention in video transformers//Proceedings of the Advances in Neural Information Processing Systems, Virtual Event, 2021: 12493-12506

[70] Liu Z, Ning J, Cao Y, et al. Video swin transformer// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual Event, 2022: 3192-3201

[71] Zolfaghari M, Singh K, Brox T. ECO: Efficient convolutional network for online video understanding//Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 713-730

[72] Diba A, Fayyaz M, Sharma V, et al. Spatio-temporal channel correlation networks for action classification//Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 299-315

[73] Wang L, Li W, Li W, et al. Appearance-and-relation networks for video classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1430-1439

[74] Zhu X, Xu C, Hui L, et al. Approximated bilinear modules for

temporal modeling//Proceedings of the International Conference on Computer Vision, Seoul, Korea, 2019: 3493-3502

[75] Kataoka H, Wakamiya T, Hara K, et al. Would mega-scale datasets further enhance spatiotemporal 3D CNNs? arXiv preprint arXiv:2004.04968

[76] Li R, Zhang Y, Qiu Z, et al. Motion-focused contrastive learning of video representations//Proceedings of the International Conference on Computer Vision, Virtual Event, 2021: 2085-2094

ZHANG Bing-Bing, Ph. D., lecturer. Her current research interests include video action recognition, image classification and deep learning.

ZHANG Jian-Xin, Ph. D., professor, M.S. supervisor. His research interests include intelligent medical image analysis and image/video recognition.

LI Pei-Hua, Ph. D., professor, Ph. D. supervisor. His research interests include deep learning, image/video recognition, object detection and semantic segmentation.

Background

Video action recognition has been a popular research topic in computer vision, attracting the attention of numerous researchers in recent decades. The research in this area can be divided into two main stages. In the first stage, numerous hand-crafted descriptors were designed for spatio-temporal representations. Currently, we are in the second stage, architecture engineering, and we categorize these architectures into four groups: two-stream CNNs, 3D CNNs, compute-efficient networks, and visual Transformer-based networks. Two-stream methods use separate networks to model appearance and motion information, which are then fused at either the intermediate or output layer. 3D CNNs naturally learn spatio-temporal features directly from RGB frames, augmenting common 2D CNNs with an additional temporal dimension. However, the heavy computational burden of 3D CNNs has led to the development of many efficient networks that trade off accuracy for speed. Transformer-based networks utilize and modify the latest powerful vision transformers to jointly encode spatial and temporal information. Nevertheless, most approaches in both stages are unimodal, overlooking the semantic information contained in the action labels.

Recently, multimodal models based on the CLIP model have shown good performance in various computer vision tasks. However, research on video action recognition is still in its early stages. To address the core issues of extending CLIP model to video action recognition, namely how to model spatio-temporal information in the visual encoder and how to perform prompt learning in the language encoder to obtain more accurate language supervision, this paper proposes a language-video contrastive learning model based on spatio-temporal information auxiliary supervision.

The proposed model is inspired by our previous works on ICCV 2017, CVPR 2018, and IEEE TPMAI 2020, which are called "Is Second-order Information Helpful for Large-scale Visual Recognition?", "Towards Faster Training of Global Covariance Pooling Networks by Iterative Matrix Square Root Normalization", and "Deep CNNs Meet Global Covariance Pooling: Better Representation and Generalization" This series of works explored the use of probability statistics represented by covariance pooling for image recognition tasks. Later on, we applied these core technologies of this series of works to the field of video action recognition and achieved good results. This achievement was published at NeurIPS 2021 with the title "Temporal-adaptive Covariance Pooling Networks for Video Recognition." The method proposed in this paper explores how to use probabilistic statistics to better model spatio-temporal information in videos under the CLIP model proposed. It is a further extension of our research group's series of works to multimodal video action recognition tasks.