夏青, 刘丹, 韦慧. 基于视觉-文本自适应提示的小样本行为识别[J]. 智能计算机与应用,2026,16(8):129-136. DOI:10.20169/j.issn.2095-2163.24122602

基于视觉-文本自适应提示的小样本行为识别

夏 青, 刘 丹, 韦 慧
(上海理工大学 机器智能研究院, 上海 200093)

摘 要: 小样本行为识别(Few-Shot Action Recognition, FSAR)旨在利用少量标签样本识别新的行为类别。鉴于视觉语言预训练模型在下游任务中的成功应用,多模态 FSAR 算法备受瞩目。然而,复杂的网络结构会导致模型参数量和计算量大幅增加。针对上述问题,本文提出一种基于视觉-文本自适应提示的多模态小样本视频行为识别方法。具体来说,该方法以对比语言-图像预训练(Contrastive Language-Image Pre-training, CLIP)模型为基础,固定其编码器内部参数,在文本模块和视觉模块中分别引入自适应提示变量,以优化多模态对比学习目标。此外,通过在图像编码器中整合可学习的轻量级适配器实现时间自适应,从而对行为的时序关系进行建模。最后,通过时序 Transformer 进一步强化视觉与文本特征的匹配过程,实现有效的小样本行为识别。在 SSv2、HMDB51 和 Kinetics 三个数据集上开展实验,验证了该模型在小样本行为识别领域的有效性。

关键词: 行为识别; 小样本学习; 自适应提示; 多模态学习; 时序建模; 特征匹配
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2026)08-0129-08

Few-shot action recognition based on visual-text adaptive prompting

XIA Qing, LIU Dan, WEI Hui
(Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Few-Shot Action Recognition (FSAR) aims to recognize new action categories with a limited number of labeled samples. Given the successful application of visual language pre-trained foundation models in various downstream tasks, multimodal FSAR algorithms have attracted much attention. However, complex network structures lead to a significant increase in model parameters and computational complexity. To address these issues, this paper presents a multimodal few-shot action recognition method based on adaptive visual-text prompts using the contrastive language-image pre-training (CLIP) model. Specifically, the proposed model freezes the internal parameters of the original visual and text encoders in the CLIP model, and introduces adaptive prompt variables to optimize the multimodal contrastive learning objective. In addition, to model the temporal relationship between action frames, learnable lightweight adapters are integrated into the visual encoder for temporal adaptation. Finally, the temporal Transformer further strengthens the feature matching process between vision and text. Experiments on three challenging datasets, SSv2, HMDB51, and Kinetics, demonstrate the effectiveness of our proposed FSAR method.

Key words: action recognition; few-shot learning; adaptive prompting; multimodal learning; temporal modeling; feature matching

0 引 言

视频动作识别是视频理解研究领域中一项基础且关键的任务,旨在从连续的视频图像中对人类或其他主体的行为进行分类。近年来,深度学习技术极大地推动了众多领域的发展[1-2],其中传统视频行为识别方法的成功很大程度上依赖大规模标注数据的支持,但这类数据的获取与标注需耗费大量人力物力。小样本动作识别(Few-Shot Action Recognition, FSAR)为解决上述难题提供了新的思路[2-4]。

在小样本行为识别研究中,主流方法多采用基于度量的元学习范式,该类方法[5-7]通过学习输入动作的视觉特征空间,根据不同动作特征的区分度,选择适当的度量标准进行数据预测。然而,基于度

基金项目: 国家自然科学基金(62106026);重庆自然科学基金(cstc2021jcyj-msxm0210);中国博士后科学基金(2020M683243)。
作者简介: 夏 青(1998—),女,硕士研究生。主要研究方向:行为识别; 韦 慧(2004—),女,学士。主要研究方向:视频分类。
通信作者: 刘 丹(1990—),女,博士,讲师,硕士生导师。主要研究方向:图像序列识别,视频理解,意图识别。Email:liudan1123@usst.edu.cn。
收稿日期: 2024-12-26

量学习的方法主要依赖单一的视觉信息,单一视觉表征的学习效率不尽人意。近年来,大规模视觉-语言预训练基础模型在多种下游任务中展现出优异的泛化能力。以 CLIP[2] 为例,该类模型利用海量且广泛的互联网数据,为下游任务提供具备泛化能力的鲁棒视觉表示,降低了不同任务适配过程中的训练成本。在当前研究背景下,直接利用文本-视频对数据进行预训练的成本相对较高,这不仅涉及大规模视频标注数据的收集与处理,还需要大量计算资源进行模型训练。因此,探索有效的知识迁移策略,将已有的图像-文本预训练模型应用于小样本视频行为识别任务,成为了一个切实可行的解决方案。

在小样本学习场景中,采用“预训练+微调”方式大量微调参数时,模型很容易出现过拟合现象,甚至陷入因大量参数更新导致的灾难性遗忘困境。在自然语言处理(Natural Language Processing, NLP)领域首次成功实施参数高效微调之后,许多研究者受到启发,开始探索在保持大型多模态预训练基础模型主体结构不变的同时,仅对一小部分额外参数进行微调,以达到适应特定下游任务的效果。当前预训练大模型大多数以图像为中心,面对视频处理任务时,AIM[8] 和 Vita-CLIP[9] 研究将少量可训练参数整合到大型基础模型中,这些方法已被证明能有效捕捉视频中的时间信息。

基于上述观察,为解决小样本学习范式中大模型微调时大量参数带来的计算压力和过拟合困境,并缓解由单一模态数据泛化能力不足的问题。本文提出了一种新颖而有效的多模态提示自适应框架,该框架依托于图像-文本对比模型 CLIP,借助语义-视觉微调技术优化模型在小样本行为识别领域的表现,并且通过整合视觉信息与文本数据来提高模型的泛化能力和学习效率,并引入自适应提示模块来优化视频和文本的联合表示,对单模态信息进行有效补充。大量实验证明了该模型的有效性,在三个广泛使用的视频行为识别相关数据集上获得了识别精度的提升,同时有效减少了训练参数。

1 相关工作

1.1 小样本学习

在小样本学习场景中,通常利用元学习技术来提升模型性能,让机器在训练样本有限的情况下,通过“学会学习”快速适应并认识新的类别[10]。当前大多数方法可分为:基于度量的方法、基于增强的方法和基于优化的方法。基于度量的方法是利用嵌入模块和度量模块来开发一个共同的特征嵌入空间,根据多种相似性度量计算未标记样本与已知类别之间的距离。文献[6]设计的原型网络和文献[11]提出的匹配网络分别采用欧几里得距离[12] 和余弦距离[13] 来获得分类结果。基于增强的技术用于直接扩大样本规模,以缓解小样本学习中数据匮乏所带来的数据压力。2019 年,文献[14]使用对抗训练生成额外的图像样本,作为缓解数据稀缺问题的解决方案。基于优化的方法侧重于学习参数优化策略或寻求更优的模型初始状态,MAML[15]、Reptile[16] 均为此类方法的典型代表,这使得模型在数据量稀缺的情况下也能够快速泛化到新任务。

1.2 小样本行为识别

小样本行为识别的核心在于小样本学习,同时也是小样本图像识别的扩展。FSAR 的最新进展可以分为两大类:单模态方法和多模态方法。

单模态方法:众多针对 FSAR 的单模态方法聚焦于挖掘强有力的特征表示。文献[7]提出的有序时间对齐网络,使用动态规划和设计时序对齐模块,对齐查询视频样本与支持集视频样本。此外,有些研究者认为单个原型并不能良好地表示视频动作的复杂性,文献[17]建议使用一种复合原型来比较视频样本相似度。文献[5]则利用 Cross-Transformer[18] 突出查询序列对应的类别原型,并通过大量不同数量帧的元组对视频进行匹配。

多模态方法:当前的多模态技术通常结合额外的模态以提高分类准确性,如运动[19]、深度[20] 和文本[21]。如文献[20]通过利用深度信息增强视觉特征,以实现高效分类。近来,视觉-语言预训练基础模型在各种下游任务中显示出惊人的潜力,文献[21]引入 CLIP 来编码语义和视觉特征,计算两种模态的原型表示并通过加权平均将其融合以增强类别原型学习。此外,CLIP-FSAR[22] 通过采用 CLIP 预训练和图像-文本对比学习目标以匹配查询样本,在小样本行为识别中取得了令人鼓舞的进展。

1.3 参数高效微调

在自然语言处理领域,参数高效微调(Parameter-Efficient Fine-Tuning,PEFT)策略已被验证,在减少迁移学习过程中,该方法可减少训练参数数量的同时,还能维持预训练模型的可靠性能。目前,在计算机视觉领域存在两大类 PEFT 模型:基于适配器(adapter)和基于提示(prompting)的模型。提示是指利用输入指令来引导模型产生期望的输

出[23]，该方法通过在模型输入端嵌入少量可训练的标记，为预训练模型适应下游任务提供了一种高效手段。文献[9]为视频任务设计了多层次提示架构，并在多个数据集上取得了令人印象深刻的结果。基于适配器的方法通常涉及在网络层间插入轻量级模块，仅在训练过程中更新这些模块。例如 ST-Adapter[24]，引入轻量级适配器，提高了原始模型的可迁移性。本文延续参数高效微调的思想，将适配器和连续提示相结合以适应于高效的视频理解任务，仅对自由变量进行更新参数，习得差异化知识，使模型在推理过程中能准确预测结果。

2　本文方法

2.1　问题设置

小样本行为识别的目的是学习一个泛化性能好的模型，对于未见过的新类别，该模型仅使用稀缺标注数据就能够精确地对行为类别进行识别。在小样本学习设定中存在两个部分：一部分是用于训练的基础数据集 $D_{\text{base}}=\{(v_i,y_i),y_i\in C_{\text{base}}\}$；另一个是用于测试的新颖数据集 $D_{\text{novel}}=\{(v_i,y_i),y_i\in C_{\text{novel}}\}$。$D_{\text{base}}$ 与 D_{novel} 的行为类别种类不存在相同的类别，即 $C_{\text{base}}\cap C_{\text{novel}}=\varnothing$。通常基于元学习的少样本行为识别任务采取以子任务为单位进行情景训练，每个子任务遵循 N-way、K-shot 范式。在训练阶段，在基础训练集中随机选择 N 个类别，再从每个类别数据中抽取 K 个样本来构成用于模型训练的支持集 $S=\{s_1,s_2,\cdots,s_{N\times K}\}$，同时从 N 类的剩余样本中抽取一部分样本作为用于模型预测的查询集 $Q=\{q_1,q_2,\cdots,q_i\}$。小样本行为识别的核心任务是在支持集中的 N 个已知类别中准确地将查询集中的样本分类。

2.2　模型框架

本文的目标是将基于视觉-语义模型有效迁移到视频域中，并应用于小样本的行为识别，以大规模预训练图像-文本对比学习模型为基础，开发自适应文本-视觉提示模型。框架如图 1 所示。

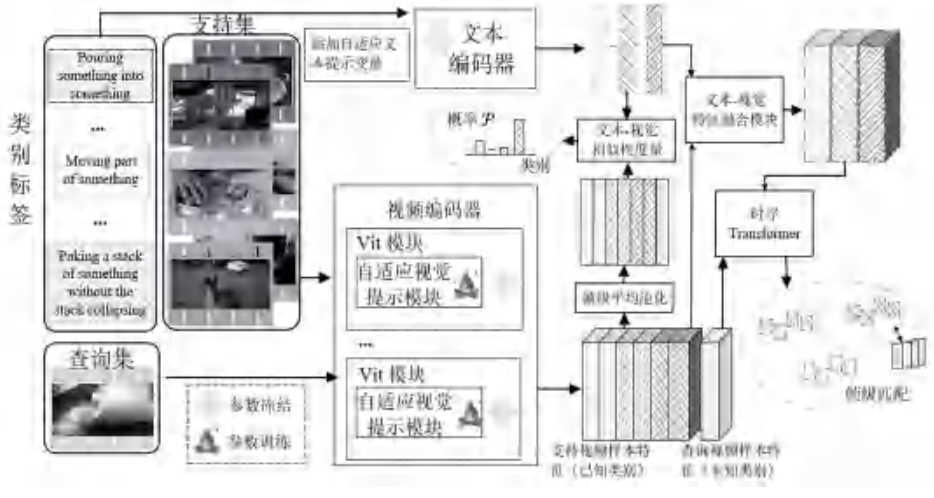


图 1　自适应多模态提示的少样本视频行为识别框架

Fig. 1　Adaptive Multimodal Prompting for Few-Shot Video Action Recognition

首先从支持集和查询集通过改制过的 CLIP 视觉编码器提取空间特征，同时使用文本编码器提取支持样本标签与自适应文本提示变量混合的语义特征；接着通过语义特征与视觉特征进行度量学习，拉近相同类别特征的度量空间；最后，视觉特征通过模态融合学习与时序增强模块以产生契合支持集的信息原型，进而与查询样本匹配，获得预测类别。

设定一个普通场景，其中 $N=3,K=2$。首先要对输入的视频数据集进行稀疏帧采样[25]以减少数据冗余信息，达到减轻模型运算量的目的。在训练子任务中，支持集表示为 $S=\{s_1,\cdots,s_i,\cdots,s_{3\times2}\}$，其中 $s_i=\{s_i^1,\cdots,s_i^t,\cdots,s_i^T\}$，同时查询集表示为 $q=\{q^1,\cdots,q^t,\cdots,q^T\}$，这里的 T 表示稀疏采样的帧数，t 表示视频帧的时间索引。对于视频-文本对 (v,y)，将视频帧输入视频编码器 $\boldsymbol{\Phi}_{\text{video}}$ 中，对应的文本输入到文本编码器 $\boldsymbol{\Phi}_{\text{text}}$ 中，分别得到支持集与查询样本的视觉特征 $\boldsymbol{F}^V$ 与 $\boldsymbol{F}^{\text{Text}}$，为了在两种模态共同映射的特征空间中拉近双方距离，本文模拟原始对比学习的训练目标，最大化两种特征的相似性。

2.3　特征编码器

2.3.1　文本特征编码器

本文使用预训练的 CLIP 文本编码器（12 层 Transformer 网络结构）对视频的语义信息生成自适应表达。具体如图 2 右侧所示，标签文本首先被转

化为可被 Transformer 编码器处理的词序列,接着在文本标记前后各添加一系列连续随机向量,也就是视频自适应提示向量。

这一系列提示向量完全由可学习自由参数组成,在训练过程中冻结文本编码器的权重,仅更新标签前后的提示向量,并通过反向传播梯度来持续优化文本周围的提示向量,使其能自适应学习视频样本相关的语义编码。这些可学习的向量构建的辅助信息,包含视频分类所需的空间时序信息。在视频-文本对 (v,y) 中,y_t 是动作标签 y 的文本格式,如 y_t = "Pouring something into something"。使用 CLIP 文本分词器生成分词后的文本序列:

$$\boldsymbol{F}_c^{\text{Text}} = \boldsymbol{\Phi}_{\text{text}}(\boldsymbol{v}_1,\cdots,\text{Tokenizer}(y_t),\cdots,\boldsymbol{v}_k) \quad (1)$$

其中,$\boldsymbol{v}_i \in \mathbb{R}^D$ 代表第 i 个提示向量,由 16 个可学习的参数组成,D 为向量维度。采用冻结的 CLIP 文本编码器 $\boldsymbol{\Phi}_{\text{text}}$ 获取 $\boldsymbol{F}_c^{\text{Text}}$,即第 C 个动作类别的语义特征。

2.3.2 视觉特征编码器

本文采用 CLIP 中的 ViT(Vision Transformer)架构[26]作为视觉特征编码器,并在该架构中补充了时空提示适应模块,以实现参数高效微调的迁移学习,使模型具备多模态学习能力。如图 2 左侧所示,对于预训练编码器采取部分冻结参数的操作,同时保持时空自适应模块的参数更新,弥补了将缺失时间维度的图像预处理模型直接用于视频域的缺陷,从而增强模型对视频动作的识别能力。

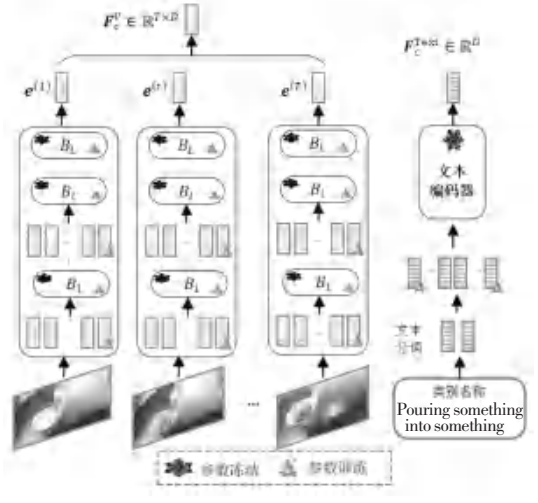


图 2 视觉-语义特征编码器

Fig. 2 Visual-Semantic Feature Encoder

在视频帧采样后,支持样本表示为 $s_i \in \mathbb{R}^{T\times H\times W\times 3}$,其中单个的待查询视频可以表示为 $q \in \mathbb{R}^{T\times H\times W\times 3}$。这里的 H 和 W 分别代表视频帧的高度和宽度,输入的通道数为 3。本文遵循 CLIP 图像编码器的处理方法,首先,一个视频样本抽取的帧依据 ViT 对图像的划分要求将其分割成 $N = \dfrac{H\times W}{P^2}$ 个大小为 $P\times P$ 的图像块。将这些图像块展平并将其投影到一系列块令牌 $\boldsymbol{Z}_t = [z_1,\cdots z_n,\cdots,z_N]$,$\boldsymbol{x}_n \in \mathbb{R}^{3\times P\times P}$,其中 $t \in \{1,2,\cdots T\}$。这些块令牌随后被转化为 D 维度的序列嵌入。然后通过逐元素相加的方式将位置信息嵌入加强到块序列中,并与一个可训练的类别令牌连接,该训练令牌由与图像令牌维度一致的自由变量构成,有助于解决自适应视频学习中由于参数冻结而丢失信息的问题。随后将 $N+1$ 个令牌送入 Transformer 块结构中。为促进帧级时序学习,图像序列经参数高效微调策略改造过后的 Transformer 块,即图 3 中改造后的 B_l 块,$l \in \{1,2,\cdots,L\}$,经视觉编码器处理后的每一帧表示为 $\{\boldsymbol{e}^{(1)},\boldsymbol{e}^{(2)},\cdots,\boldsymbol{e}^{(T)}\} \in \mathbb{R}^{T\times D}$,最终获得具有时序信息的 C 类别样本视频特征 $\boldsymbol{F}_c^V \in R^{T\times D}$,对视频编码器内部结构的修改详情可见 2.4 节。

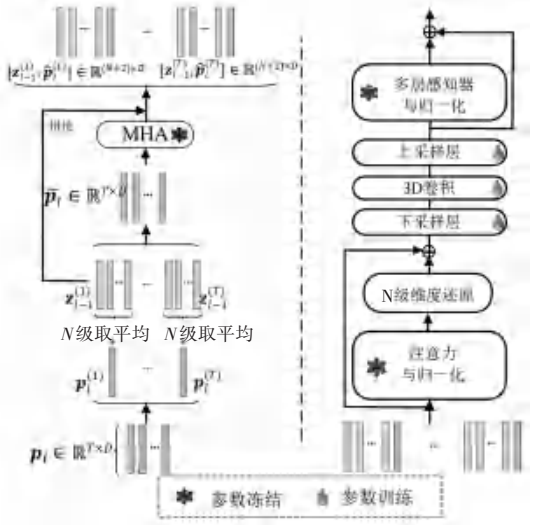


图 3 时空提示适应器

Fig. 3 Spatiotemporal Prompt Adapter

2.4 时空提示适应器

为引导模型高效学习视频样本的时空表征,如图 3 所示,2.3 节叙述的自适应时间视觉提示信息被策略性地合并在每个 Transformer 层的输入空间,即每一层都包含可学习的视觉变量,其表示为 $\boldsymbol{P}^{\text{Temp}} \in \mathbb{R}^{L\times T\times D}$,其中 L 是 CLIP 模型中 Transformer 的层数,将可学习的提示的集合 $\boldsymbol{P}^{\text{Temp}} = \{\boldsymbol{p}_l \in \mathbb{R}^{T\times D} \mid l = 0,\cdots,L-1\}$ 作为输入,其中 $t \in \{1,2,\cdots,T\}$。第 t 帧的嵌入与对应的时间提示连接方式如下式:

$$\widetilde{\boldsymbol{p}}_l^{(t)} = \boldsymbol{p}_l^{(t)} + \frac{1}{N+1}\sum_{j=1}^{N+1}(z_{l-1}^{(t)})_j \quad (2)$$

式中 $(z_{l-1}^{(t)})_j$ 代表特征维度上第 j 维元素。

如图 3 所示，该过程进入多头注意力机制，该机制利用所有时间提示来捕获视频内的全局时空依赖性。在第 l 块的操作被表达为：

$$\hat{\boldsymbol{p}}_l = \mathrm{MHA}(\mathrm{LN}(\tilde{\boldsymbol{p}}_l)) \tag{3}$$

其中，$\hat{\boldsymbol{p}}_l = [\tilde{\boldsymbol{p}}_l^{(1)}, \tilde{\boldsymbol{p}}_l^{(2)}, \cdots, \tilde{\boldsymbol{p}}_l^{(T)}]$；MHA 表示从 CLIP 导出并冻结参数的注意力模块；LN 表示归一化层，最后学习的时间提示与帧嵌入连接，紧接着将送入调制后的 Transformer 块中。

为进一步强化视频时序信息，在冻结参数的 Transformer 块中嵌入可学习的时空自适应器，该适应器主要由 3D 卷积模块构成，具体地说，适配器模块由一个下投影线性层、一个非线性激活函数、一个上投影线性层和一个残差连接组成。给定第 l 层的特征表示 $\boldsymbol{S}^{(l)}$，其中 $\boldsymbol{S}^{(l)} = [\boldsymbol{z}_{l-1}, \boldsymbol{p}_l] \in \mathbb{R}^{T\times(N+1)\times D}$，适配器模块的输出可以写为：

$$\boldsymbol{S}^{(l+1)} = \boldsymbol{S}^{(l)} + \mathrm{Adapter}(\boldsymbol{S}^{(l)}) \tag{4}$$

$$\mathrm{Adapter}(\boldsymbol{S}^{(l)}) = f_{\mathrm{up}}(\mathrm{Conv3D}(f_{\mathrm{down}}(\boldsymbol{S}^{(l)}))) \tag{5}$$

其中，f_{down} 为下投影层；f_{up} 为上投影层；Conv3D 表示三维卷积函数。

在网络学习视觉分支时，只微调适配器模块的参数以提高模型效率。最终，得到视觉特征 $\boldsymbol{F}^V \in \mathbb{R}^{T\times D}$，支持集特征 $\boldsymbol{F}_s^V$ 和查询特征 $\boldsymbol{F}_q^V$。

2.5　多模态时序增强模块

对于支持集，给定一个视觉基础原型 $\boldsymbol{P}^V \in \mathbb{R}^{T\times D}$，和对应的文本标签特征 $\boldsymbol{F}^{\mathrm{Text}} \in \mathbb{R}^{1\times D}$，应用交叉注意力机制获得多模态原型特征 $\boldsymbol{P}^{\mathrm{VT}}$，具体来说，查询-键值三元组 $\boldsymbol{q}^{\mathrm{VT}}$、$\boldsymbol{k}^{\mathrm{VT}}$、$\boldsymbol{v}^{\mathrm{VT}}$，按如下方式获得：

$$\boldsymbol{q}^{\mathrm{VT}} = \boldsymbol{P}^V + \mathrm{Repeat}(\boldsymbol{F}^{\mathrm{Text}}) \tag{6}$$

其中，Repeat 表示将 $\boldsymbol{F}^{\mathrm{Text}}$ 特征复制 T 次以匹配 $\boldsymbol{P}^V$ 的形状。

$$\boldsymbol{k}^{\mathrm{VT}} = \boldsymbol{v}^{\mathrm{VT}} = \mathrm{Concat}(\boldsymbol{P}^V, \boldsymbol{F}^{\mathrm{Text}}) \tag{7}$$

利用 Transformer 融合视频-文本多模态信息

$$\boldsymbol{P}^{\mathrm{VT}} = \mathrm{Transformer}(\boldsymbol{q}^{\mathrm{VT}}, \boldsymbol{k}^{\mathrm{VT}}, \boldsymbol{v}^{\mathrm{VT}}) \tag{8}$$

由于查询集中没有额外的文本信息，设置 $\boldsymbol{k}_Q^{\mathrm{VT}} = \boldsymbol{v}_Q^{\mathrm{VT}} = \boldsymbol{q}_Q^{\mathrm{VT}} = \boldsymbol{F}_Q^V$，并使用与支持集共享权重的网络来学习查询样本的增强特征以提高模型效率。经时序增强模块后可获得包含语义信息的 $\boldsymbol{P}^{\mathrm{VT}}$，查询集由于其未知类别的特性进行时序增强获得特征 $\boldsymbol{F}_Q^{V'}$。

2.6　损失函数

为弥合 CLIP 和小样本视频任务之间的任务差异，采用正样本对和负样本对的对比学习方式，将视频样本与其对应的类别语义文本构建为正样本对，视频-文本对比学习的优化目标为最大化匹配视频特征与对应文本特征之间的余弦相似度。视频-文本跨模态匹配概率定义如下：

$$\boldsymbol{p}^{\mathrm{Vt}}_{(y=c|v=s_i)} = \frac{\exp(\mathrm{sim}(\mathrm{TAP}(\boldsymbol{F}_{s_i}^V), \boldsymbol{F}_c^{\mathrm{Text}})/\boldsymbol{\tau})}{\sum_{j=1}^{N}\exp(\mathrm{sim}(\mathrm{TAP}(\boldsymbol{F}_{s_i}^V), \boldsymbol{F}_j^{\mathrm{Text}})/\boldsymbol{\tau})} \tag{9}$$

其中，$\mathrm{sim}(\cdot,\cdot)$ 表示余弦相似度函数；TAP 指的是时间平均池化；τ 表示可学习因子。

随后通过预测值和实际标签之间的交叉熵损失 $\mathcal{L}_{v2t}$ 来优化目标函数。接着根据文献[7]的度量标准，通过时间度量模块 $\mathcal{M}$ 获得支持-查询距离。将查询样本匹配到相应的类别原型获得标准交叉熵损失 $\mathcal{L}_{q2s}$，支持-查询样本距离计算公式如下：

$$D_{s,q} = \mathcal{M}(\boldsymbol{P}^{\mathrm{VT}}, \boldsymbol{F}_Q^{'}) \tag{10}$$

其中，$\boldsymbol{p}^{\mathrm{VT}}$ 为经跨模态对比学习校准后的特征分布，$\boldsymbol{F}_Q^{'}$ 代表查询样本时序特征。

综上，本文所提方法的联合总损失函数为视觉-文本对比损失与原型匹配判别损失的加权组合：

$$\mathcal{L}_{\mathrm{total}} = \lambda\,\mathcal{L}_{v2t} + \mathcal{L}_{q2s} \tag{11}$$

其中，λ 为损失平衡超参数，用于动态权衡跨模态对比学习与原型判别学习两项优化目标的训练权重，避免单一损失主导模型更新过程。

3　实验与评估

3.1　数据集

本文实验采用小样本行为识别领域广泛使用的三个公开基准数据集：Something-Something v2[27] (SSv2)、HMDB51[28] 和 Kinetics[1]。这三类数据集均广泛应用于视频理解、动作识别等任务，但 SSv2 数据集中包含不同时长的视频，有利于研究视频中的时序信息。在小样本数据集制备方面，由于 HMDB51 由 51 个动作类组成，即挑选其 51 个类别，每个类别包含 100 个视频，按照 31、10、10 个互不重叠的动作类别依次划定训练集、验证集和测试集。从 SSv2 和 Kinetics 数据集中选取 100 个类，同样每个类也包含 100 个视频，将这两个数据集按照 64、12 和 24 个动作类别进行划分。以上 3 种数据集的划分均依据于文献[7]中提出的分割方法。

3.2　实验细节

本研究构建了面向小样本行为识别的实验框架，视觉编码器主干采用在 4 亿对网络图像-文本对上完成预训练的 ViT-B/16 模型。数据预处理流程与文献[25]保持一致，从每个视频素材中均匀切

分 8 个片段,并通过随机抽样的方式从每个片段中抽取一帧作为当前时刻的视觉数据。为进一步处理这些帧,对其进行随机裁剪以生成 224×224 图像区域,在模型元训练阶段,使用 Adam 优化器优化模型,初始学习率为 0.000 05,平衡因子最初设置为 1。为保证模型性能的同时减少计算资源的消耗,混合精度训练贯穿整个学习过程。本文方法在训练和测试过程中均在 PyTorch 框架中实现,并配备一张 NVIDIA RTX A6000 显卡,显存大小为 48 GB。

3.3 评估标准

本研究在公开的小样本数据集上开展对比实验,该实验在当下两种主流评估方案“5-way 1-shot”与“5way,5-shot”中进行。该设置中,从数据集中随机选取 5 个不同的类别,每个类别仅使用 1 个样本作为支持样本,并随机抽取 1 个样本构成查询样本,用以辅助查询集预测正确类别。其中每个 episode 由这样随机抽取的支持集和查询集组合而成。类似地,“5-way,5-shot”同样选取 5 个类别,每个类别使用 5 个样本来形成支持集。一个 episode 由随机抽取的支持集和查询集组成,在推理阶段,本文通过计算 10 000 轮次后的平均分类准确率来评估模型的识别性能。同时,使用参数量(Parameters)来衡量模型参数的内存占用情况。

3.4 实验结果分析

为验证本文所提方法的有效性,表 1 展现了本方法在 3 个经典视频行为识别数据集上的结果,本文方法在 Kinetics 和 SSv2-small 数据集上实现了领先性能。在 SSv2-small 数据集上,相较于 CLIP-FSAR 方法,1-shot 设置下准确率提升了 2.4 个百分点,5-shot 设置下提升了 3.2 个百分点,在 Kinetics 数据集上,1-shot 设置下准确率提升 1 个百分点,5-shot 设置下准确率提升 0.7 个百分点。此外,提出的方法在 HMDB51 数据集的单样本任务中也获得了 0.1 个百分点的增益。本文对 SSv2-small 数据集在 5-way,1-shot 实验设定中,与其他全微调方法进行了可训练参数量对比,CLIP-FSAR 方法的参数量为 90.99 M,而本文方法仅用 24.9 M 的参数对模型进行微调,结果带来 2.4 个百分点的提升。总的来说,本文方法有效地减少了模型训练过程中的微调参数量,同时在主流数据集上的测试结果相较前人的研究也实现了精度的提升。

表 1 在 SSv2-small、Kinetics 和 HMDB51 数据集上与现有方法的分类准确率比较

Table 1 Comparison of classification accuracy with existing methods on the SSv2-small, Kinetics, and HMDB51 Datasets

方法	主干网络	参数更新	HMDB51		Kinetics		SSv2-small	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
STRM[29]	Inet-RN50	全更新	52.3	77.3	62.9	86.7	37.1	55.3
HyRSM[30]	Inet-RN50	全更新	60.3	76.0	73.7	86.1	40.6	56.1
HCL[31]	Inet-RN50	全更新	59.1	73.3	73.7	85.8	38.7	55.4
Huang et al[17]	Inet-RN50	全更新	60.1	77.0	73.3	86.4	38.9	61.6
AMeFu-Net[20]	Inet-RN50	全更新	60.2	75.5	74.1	86.8		
SRPN[21]	Inet-RN50	全更新	61.6	76.2	75.2	87.1		
MoLo[19]	Inet-RN50	全更新	60.8	77.4	74.0	85.6	41.9	56.2
CLIP-FSAR[22]	CLIP-ViT-B	全更新	75.8	**87.7**	89.7	95.0	54.5	61.8
本文	CLIP-ViT-B	部分更新	**75.9**	84.5	**90.7**	**95.7**	**56.9**	**65.0**

4 消融实验

4.1 提示的消融

为全面评估自适应提示在不同模态下对模型性能的具体影响,本文设计了一系列消融实验,具体汇总于表 2。在未使用任何语义提示的情况下,模型仅依赖视觉信息进行识别,其在 SSv2-small 和 Kinetics 数据集上的准确率分别为 46.8%、84.4%。这一结果作为实验基准,用于衡量后续加入视觉-语义提示后的性能提升。模型中引入自适应时空提示后,可观察到在 SSv2-small 数据集上的准确率提升到了 54.7%,Kinetics 数据集上提升至 89.3%的准确率。与没有自适应时空提示的情况相比,仅仅依赖于语义时序提示的方法在 SSv2-small 和 Kinetics 数据集上表现略逊色于仅依赖于自适应时空提示的模型。此外,当模型同时使用视觉、文本模态时,在这两个具有挑战性的数据集上的识别性能进一步提升到 56.9%和 90.7%。这些结果表明本文设计的提示向量能够丰富视觉表征,有效补充了仅用动作类别名称所获得的信息。此外,实验结果还揭示了

双模态提示向量相较于单模态提示向量在性能上的显著优势。这种优势可能源于双模态提示向量能够更全面地捕捉视频行为的语义信息，从而在识别过程中提供更为丰富的上下文支持。

表 2　在 SSv2-small 和 Kinetics 数据集上的模块消融实验

Table 2　Ablation study on the modules within the SSv2-small and Kinetics datasets

模块方法选择		数据集准确率/%	
语义自适应提示	时空自适应提示	SSv2-small	Kinetics
×	×	46.8	84.4
×	✓	54.7	89.3
✓	×	47.1	89.6
✓	✓	56.9	90.7

4.2　平衡因子的消融

本文在 SSv2-small 和 HMDB51 数据集上开展消融实验，探究式(11)中超参数 λ 对模型性能的影响，如图 4 所示，当超参数 λ 设定为 1.5 时，模型在两个数据集“5-way，1-shot”的设定下的表现分别为 56.9%与 78.5%。当不涉及前述视频-文本对比损失时，模型仅通过视频原型匹配损失进行优化，即便引入视频标签提示与时序增强模块，效果仍然欠佳。随着 λ 数值的增大，视频分类精度得到了显著提升。这一趋势表明，缩短视觉特征与标签表示之间的距离，有益于在小样本行为识别任务中提取决定分类的关键特征。然而，随着数值继续增大，模型在不同数据集上的精度曲线不再保持上升状态，而是趋于稳定，甚至出现下降的情况。这一结果表示，本方法为小样本行为识别任务带来的性能提升超过某个阈值后会达到饱和，两种损失需要达到平衡，模型的整体性能才能达到最大化。

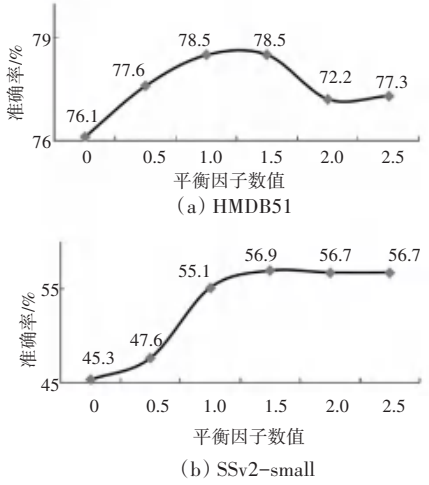


图 4　平衡因子变化曲线图

Fig. 4　Balance factor variation curve

5　结束语

本研究基于 CLIP 模型进行扩展，提出一种简单而高效的自适应多模态提示模型，该模型适用于小样本视频行为识别任务。在模型的语义与视觉编码器中添加应用混合自适应提示策略，在语言分支上，强化视频域的语言表达，自适应生成每个类的语义提示。在视觉分支上构建时空联合提示自适应模块，通过 Adapter 与 Prompt 相结合建模视频连续帧之间的时间依赖信息，完成模型由图像域到视频域的时序学习。此外，在 3 个具有挑战性的数据集上的准确率表明，本文方法在降低训练参数量的同时，准确率也得到了明显的提升，尤其是在强调时间相关的数据集 SSv2-small 上有了显著的提高。

参考文献

[1] Carreira J, Zisserman A, Quo V. Action recognition? A new model and the Kinetics dataset [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6299.

[2] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]// Proceedings of the International Conference on Machine Learning. ICML, 2021: 8748.

[3] 赵凯琳, 靳小龙, 王元卓. 小样本学习研究综述[J]. 软件学报, 2021, 32(2): 349-369.

[4] 包震伟, 刘丹, 米金鹏. 弱监督与小样本学习场景下视频行为识别综述[J]. 计算机应用研究, 2023, 40(6): 1629.

[5] Perrett T, Masullo A, Burghardt T, et al. Temporal-relational crosstransformers for few-shot action recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 475.

[6] Snell J, Swersky K, Zemel R S. Prototypical networks for few-shot learning [C]// Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, 2021: 2644.

[7] Cao K, Ji J, Cao Z, et al. Few-shot video classification via temporal alignment[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10618.

[8] Yang T, Zhu Y, Xie Y, et al. Aim: Adapting image models for efficient video action recognition [EB/OL]. (2023-02-06) [2024-12-09]. https://doi.org/10.48550/arXiv.2302.03024.

[9] Wasim S T, Naseer M, Khan S, et al. Vita-clip: Video and text adaptive clip via multimodal prompting [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 23034.

[10] 蔡奇, 李凡长. 类特征增强元学习算法[J]. 小型微型计算机系统, 2022, 43(2): 225.

[11] Vinyals O, Blundell C, Lillicrap T, et al. Matching networks for one shot learning [C]//Proceedings of the Advances in Neural Information Processing Systems. 2016: 1049.

[12] Wang Y, Chao W L, Weinberger K Q, et al. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning

[EB/OL]. (2019-11-12)[2024-12-09]. https://doi.org/10.48550/arXiv.1911.04623.

[13] Qiao S, Liu C, Shen W, et al. Few-shot image recognition by predicting parameters from activations[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7229.

[14] Pahde F, Ostapenko O, Hnichen P J, et al. Self-paced adversarial training for multimodal few-shot learning[C]//Proceedings of IEEE Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE, 2019: 218.

[15] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [C]//Proceedings of International Conference on Machine Learning. New York: ACM, 2017: 1126.

[16] Nichol A. On first-order meta-learning algorithms[EB/OL]. (2018-03-11)[2024-12-09]. https://doi.org/10.48550/arXiv.1803.02999.

[17] Huang Y, Yang L, Sato Y. Compound prototype matching for few-shot action recognition [C]//Proceedings of European Conference on Computer Vision. Piscataway, NJ: IEEE, 2022: 351.

[18] Doersch C, Gupta A, Zisserman A. Crosstransformers: Spatially-aware few-shot transfer[C]//Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, 2020: 21981.

[19] Wang X, Zhang S, Qing Z, et al. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 18011.

[20] Fu Y, Zhang L, Wang J, et al. Depth guided adaptive meta-fusion network for few-shot video recognition[C]// Proceedings of 8th ACM International Conference on Multimedia. New York: ACM, 2020: 1142.

[21] Wang X, Ye W, Qi Z, et al. Semantic-guided relation propagation networkfor few-shot action recognition [C]// Proceedings of 29th ACM International Conference on Multimedia. New York: ACM, 2021: 816.

[22] Wang X, Zhang S, Cen J, et al. CLIP-guided prototype modulating for few-shot action recognition[J]. International Journal of Computer Vision, 2024, 132(6): 1899.

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

[24] Pan J, Lin Z, Zhu X, et al. St-adapter: Parameter-efficient image-to-video transfer learning [C]//Proceedings of the Advances in Neural Information Processing Systems. NeurIPS, 2022: 26462.

[25] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks for action recognition in videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 37(3): 2740.

[26] Dosovitskiy A. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22)[2024-12-09]. https://doi.org/10.48550/arXiv.2010.11929.

[27] Goyal R, Ebrahimi K S, Michalski V, et al. The "something something" video database for learning and evaluating visual common sense [C]//Proceedings of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 5842.

[28] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition[C]// Proceedings of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2011: 2556.

[29] Thatipelli A, Narayan S, Khan S, et al. Spatio-temporal relation modeling for few-shot action recognition[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 19958-19967.

[30] Wang X, Zhang S, Qing Z, et al. Hybrid relation guided set matching for few-shot action recognition[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 19948.

[31] Zheng S, Chen S, Jin Q. Few-shot action recognition with hierarchical matching and contrastive learning[C]// Proceedings of European Conference on Computer Vision. Piscataway, NJ: IEEE, 2022: 297.