doi：10. 12068/j. issn. 1005－3026. 2026. 20250040

基于跨模态注意力机制的视频-文本检索方法

董 闯[1]，栗 伟[1,2]，巴 聪[1]，覃文军[1,2]
（1. 东北大学 计算机科学与工程学院，辽宁 沈阳 110819；
2. 东北大学 工业智能与系统优化国家级前沿科学中心，辽宁 沈阳 110819）

摘 要：针对当前视频-文本检索方法未能有效结合时间信息与相关性信息进行联合建模的问题，提出一种基于跨模态注意力机制的视频-文本检索方法．首先，利用预训练的大规模图像-文本模型提取文本和视频帧的嵌入表示，通过知识迁移缓解不同模态数据之间的异质性问题．然后，使用联合文本-帧跨模态注意力机制模块，同时编码视频帧之间的时间信息以及视频帧与文本之间的相关性信息，捕获更具竞争力的视频特征表示．最后，利用交叉熵损失函数约束模型训练．通过对比实验验证，该方法能够有效捕获视频帧的时间信息和相关性信息，在MSR-VTT(microsoft research video to text)和LSMDC(large-scale movie description challenge)数据集上取得具有竞争力的效果．

关 键 词：视频-文本检索；跨模态；注意力机制；知识迁移；视频特征表示
中图分类号：TP 391 **文献标志码**：A **文章编号**：1005-3026(2026)01-0075-07

Video-Text Retrieval Method Based on Cross-Modal Attention Mechanism

DONG Chuang[1], *LI Wei*[1,2], *BA Cong*[1], *TAN Wen-jun*[1,2]
（1. School of Computer Science & Engineering, Northeastern University, Shenyang 110819, China; 2. National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, Shenyang 110819, China. Corresponding author: LI Wei, E-mail: liwei@cse.neu.edu.cn）

Abstract: Existing video-text retrieval methods fail to effectively model temporal information and relevance information in a unified manner. To address this issue, a video-text retrieval method based on a cross-modal attention mechanism was proposed. Firstly, embeddings of video frames and text were extracted using a large-scale pre-trained image-text model, and knowledge transfer was leveraged to alleviate the heterogeneity between different modalities. Then, a joint text-frame cross-modal attention module was introduced to simultaneously encode temporal information among video frames and relevance information between video frames and text, enabling the capture of more competitive video representations. Finally, the cross-entropy loss function was used to constrain the model training. Comparative experiments for verification demonstrate that the proposed method can effectively capture temporal and relevance information of video frames, achieving competitive performance on the microsoft research video to text (MSR-VTT) and large-scale movie description challenge (LSMDC) datasets.

Key words: video-text retrieval; cross-modal; attention mechanism; knowledge transfer; video representation

视频-文本检索(video-text retrieval,VTR)是一项多模态任务，通过输入文本信息查询系统需要返回数据库中与之对应的最相似的视频内容，输入视频查询返回最相似的文本.VTR技术的应用使得用户能够快速准确地查找所需信息，并更好地理解海量视频数据，在短视频平台、视频网站等平台大量应用.一种常见的方法是将文本和视频特征嵌入到一个共享空间中，然后通过度量

收稿日期：2025-04-22
基金项目：高等学校学科创新引智计划项目(B16009).
作者简介：董 闯(1994—)，男，辽宁本溪人，东北大学博士研究生.
通信作者：栗 伟，E-mail:liwei@cse.neu.edu.cn.

它们之间的相似性来实现匹配.然而,文本和视频之间存在异质性,即不同模态数据之间的嵌入表示不能直接比较,使得在同一空间度量2种模态数据的相似性变得极具挑战.

最近,开创性的语言-图像预训练[1](contrastive language-image pre-training,CLIP)模型,通过学习4亿个图像-文本对的通用视觉和语言表示,在多个视觉语言任务中取得了具有竞争力的表现.一些研究工作[2-3]将CLIP模型从大规模数据中学习到的图像语言知识应用于VTR中,这种迁移能够在一定程度上弥补视觉和语言之间的模态差异.然而,视频相对于图像具有更为丰富的多模态和时间信息.因此,如何有效地将CLIP中学到的图像知识迁移到视频领域对于VTR的发展至关重要.

当前的VTR模型在将图像知识迁移到视频时,主要关注时间信息和相关性信息的建模.研究者已经尝试探索不同的时间建模方法,包括后验结构[2,4-5]、中间结构[6]和分支结构[7]等.然而,这些方法并未充分考虑视频中均匀采样的帧与视频主要内容之间的相关性.如图1a所示,按时间顺序展示了从MSR-VTT[8]数据集中采样的多个视频帧.视频的主要内容是"A man is driving a car".不同帧所传达的信息可能与视频的整体内容有不同程度的相关性,甚至可能传达误导性信息.若聚合这些帧而不考虑其具体特征,将导致次优的视频特征.此外,视频帧的顺序信息同样会影响视频特征表示的准确性.如图1b所示,顺序和倒序所传递的信息完全是相反的,但现有工作[9]则忽略了时间信息的建模.因此,同时建模时间信息和相关性信息对于将图像知识迁移到视频领域至关重要.

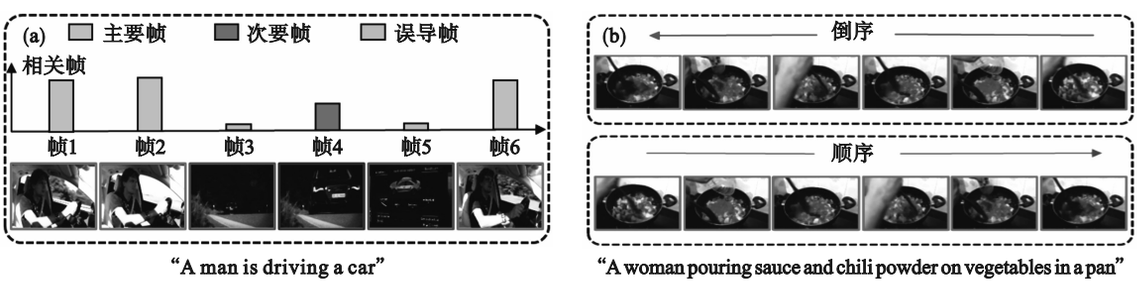

图1 图像知识迁移到视频过程中相关性信息和时间信息的影响

Fig. 1 Impact of relevance information and temporal information during process of transferring image knowledge to videos

(a)—相关性信息;(b)—时间信息.

为解决上述问题,同时建模时间信息和相关性信息,本文提出基于跨模态注意力机制的VTR方法.首先,将预训练的CLIP模型中的跨模态知识迁移到视频-文本中;其次,设计2种不同的跨模态注意力机制提升VTR的性能,分别为基于TopK文本-帧跨模态注意力机制的VTR模型,以及在其基础上改进的基于联合文本-帧跨模态注意力机制的VTR模型.最后,在多个数据集上验证并分析该方法的有效性.

1 相关工作

受限于视频文本数据集的数量,早期的VTR工作[10-12]从"专家"[13]模型中预先提取视频中的多模态特征,再设计复杂的融合机制将多模态特征融合为最终的视频表示.HowTo100M[14]和WebVid-2M[15]等大规模视频数据集的可用性为训练模型学习视频及其对应文本的有效表示提供了丰富的数据.大规模数据集的出现推动了基于预训练微调范式的视频检索模型[16]的快速发展.然而,这一范式在预训练阶段需要大量的计算资源,制约了这类方法的发展.

最近,基于CLIP[1]的视频-文本检索模型取得了显著的成功[2-5,7,9],将CLIP模型学习到的图像-文本知识扩展到视频,能够有效缓解不同模态数据之间的异质性鸿沟.CLIP从大量的图像-文本对中学习通用的视觉语言表示,这在多模态任务中显示出显著的竞争力.将图像知识转化为视频的关键在于对视频的时间信息和相关性信息进行建模.获取时间信息的模型已经被广泛研究,CLIP4Clip[2](CLIP for video clip retrieval)采用LSTM[17](long short-term memory)和加入位置编码的Transformer[18]来建模时间信息.STAN[7](spatial-temporal auxiliary network)采用分支结

构,分解了时空模块,可以更好地理解视频中的视觉和时间关系.此外,Bertasius等[6]也证明捕获时间信息可以增强视频表示.这些方法忽略了视频帧与主要内容之间的相关性,另一部分研究者专注于相关性信息的建模.ATP[19](atemporal probe)模型表明,对时间信息的深刻理解并不总是实现强大或最先进性能所必需的.X-Pool[9]设计了一种跨模态注意力机制来衡量不同帧的权重.区别于以上研究,本文模型设计了一个更精细的架构,将图像知识转移到视频中,可以同时建模多帧之间的时间和相关性信息.

2　方　　法

本章将介绍本文的VTR方法的主要构成,模型的主要架构如图2所示.模型中视频帧编码器和文本特征编码器用于编码视频帧和文本的特征表示;文本-帧跨模态注意力机制模块用于交互视频帧和文本的信息,得到最终的视频特征表示.最后计算视频和文本的相似性,使用损失函数反向约束视频和文本编码器的编码.

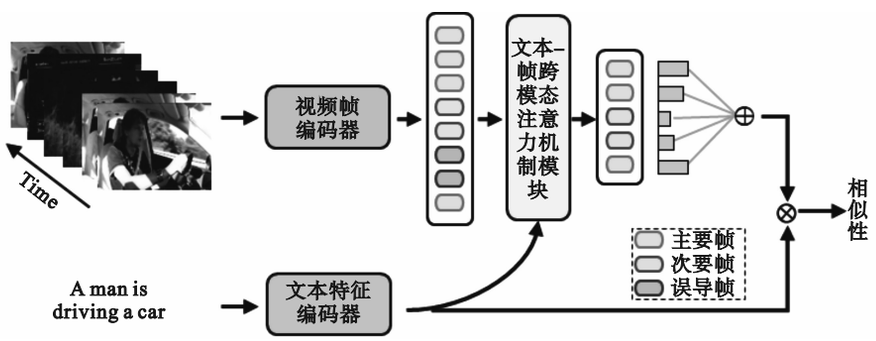

图2　基于跨模态注意力机制的视频-文本检索模型架构

Fig. 2　Architecture of video-text retrieval model based on cross-modal attention mechanism

2.1　视频文本特征提取

在最近的研究中,利用预训练的图像-文本模型来指导视频-文本模型的训练已被证明是可行和有效的.通过将图像-文本知识转移到视频-文本领域,减少了模型对大量视频数据的依赖,节约了计算资源.为了与近期的研究[2]进行公平比较,模型采用CLIP[1]作为主干框架,提取文本特征和多帧特征.本文专注于探究如何聚合多个视频帧,充分捕获多帧数据之间的时间信息和相关性信息,获取更具代表性的视频特征表示,以提高检索的准确率.

VTR的目标是学习视频v和文本t之间的相似性函数$s(v,t)$,使相关的视频文本对相似性变高,不相关的视频文本对相似性变低.具体来说,模型的输入包括文本数据$\boldsymbol{T}$和视频数据$\boldsymbol{V}$.其中,$\boldsymbol{T}$表示一个$b\times l$的矩阵,b表示批大小,l表示文本序列的长度;$\boldsymbol{V}$矩阵的形状为$b\times n\times c\times h\times w$,$n$是视频帧的数量,$c$是通道数,$h$,$w$是图像的高和宽.使用预训练的CLIP模型编码图像和文本特征,得到每个视频文本对的文本嵌入表示t,f表示了帧的嵌入表示,以及n帧的嵌入表示$\boldsymbol{v}=\left[f_1,f_2,\cdots,f_n\right]^{\mathrm{T}}$.视频文本编码器采用CLIP模型的ViT-B/32结构.获取视频特征表示的关键是如何聚合多个视频帧特征,需要充分考虑视频帧的时间信息以及相关性信息.

2.2　TopK文本-帧跨模态注意力机制

首先讨论TopK文本-帧跨模态注意力机制对模型的影响.均匀采样的视频帧与视频内容之间的相关性存在差异,甚至包含误导信息的帧被采样用于训练.受到ATP[19]模型的启发,图像级别的理解与完整视频级别的理解相比同样具备竞争力.然而,单帧图像特征表示整个视频内容,稳定性较差,特别是对于场景变化频繁的视频.如图3所示,本文设计TopK文本-帧跨模态注意力机制,消除误导帧对模型的影响.

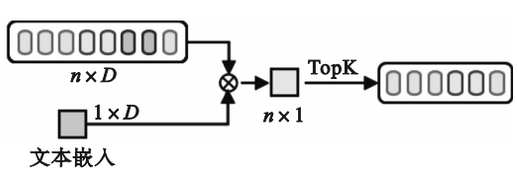

图3　TopK文本-帧跨模态注意力机制

Fig. 3　TopK text-frame cross-modal attention mechanism

给定编码器编码后维度为D的n个帧的特征表示以及文本特征;TopK跨模态注意力机制的目标是返回最能代表视频主要内容的前K个帧.为此,计算文本特征与n个帧特征之间的余弦相

似性,去除误导性帧对模型的影响.对于每个视频文本对,这一过程可以表示为

$$v_{\text{attention}}=\text{TopKAttention}\left(v|t\right). \tag{1}$$

其中:TopKAttention 表示 TopK 跨模态注意力机制;$v_{\text{attention}}$ 表示经过注意力后的视频特征.

经过 TopK 跨模态注意力后,得到最能代表视频内容的 K 个帧.对于 K 的选择,选择 1 个太大的值可能不能完全消除误导帧,而选择 1 个太小的值可能会导致视频信息的丢失.在消融实验中,本文将讨论 K 对模型准确率的影响.对于不同的数据,K 的选择将会有所差异,2.3 节将讨论一种自适应加权的跨模态注意力机制,以缓解 K 选择问题.

2.3 联合文本-帧跨模态注意力机制

针对 TopK 注意力机制的局限,设计一种联合注意力机制方法.不同于 TopK 生硬地将不相关的帧去除而仅保留相关的帧,本文提出的联合文本-帧跨模态注意力机制(见图 4)能够同时建模帧之间的时间信息和相关性信息,并且自适应地为每个帧分配适当的权重,能够应对不同视频数据,无需手动设置参数.

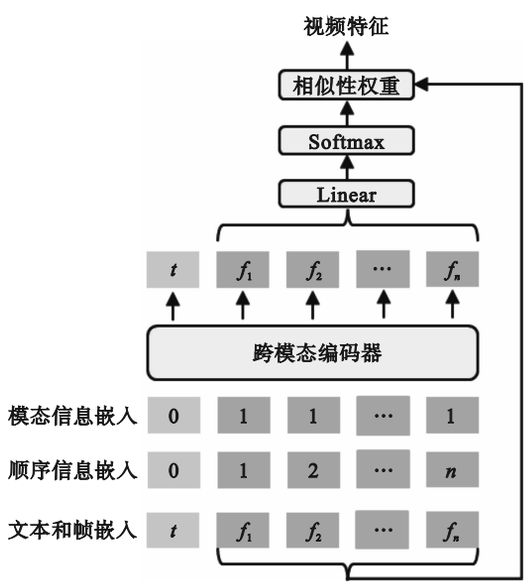

图 4 联合文本-帧跨模态注意力机制

Fig. 4 Joint text-frame cross-modal attention mechanism

图 4 为联合文本-帧跨模态注意力机制的主要构成.将文本嵌入表示与视频帧的嵌入表示连接在一起,组成模型的基本输入,将可学习的时间嵌入表示以及模型信息加入初始的视频文本数据中,作为联合跨模态编码器的输入.随后将数据输入到跨模态编码器中,其基本结构由 Transformer[18] 编码器构成.该模块可以实现帧与文本之间的细粒度信息交互,同时可以实现帧之间的交互.编码器的结果通过 Linear 层以及 Softmax 函数计算出视频帧与文本数据的相似性权重.最后,利用得到的相似性权重对原始帧数据进行加权,得到完整的视频特征表示.这一过程可以表示为

$$v_{\text{attention}}=\text{JointAttention}\left(v|t\right). \tag{2}$$

其中:JointAttention 表示联合文本-帧跨模态注意力机制.最终的视频特征表示 $v_{\text{attention}}$ 由 t 条件下的 v 的加权聚合而成.

2.4 损失函数

在视频文本检索任务中,为了最大化配对视频与文本之间的相似性,同时最小化不匹配对的相似性,本文采用对称交叉熵损失对模型进行优化.给定一个批量大小为 b 的视频文本对,构建一个 $b\times b$ 的相似性矩阵.矩阵对角线元素表示配对的视频和文本之间的相似性,其他位置元素表示不配对的视频和文本之间的相似性.训练目标是使矩阵对角线上匹配对的相似性最大化,同时抑制非对角线上的非匹配对.具体地,损失函数包括 2 个方向,视频到文本检索损失(l_{v2t}):

$$l_{v2t}=-\frac{1}{b}\sum_{i=1}^{b}\log\frac{\exp\left(s\left(v_i,t_i\right)\right)}{\sum_{j=1}^{b}\exp\left(s\left(v_i,t_j\right)\right)}. \tag{3}$$

其中:$s\left(v_i,t_i\right)$ 表示文本 t_i 与视频 v_i 之间的相似性.该损失鼓励在给定视频的情况下,将正确匹配的文本排在所有文本中的首位.文本到视频检索损失(l_{t2v}):

$$l_{t2v}=-\frac{1}{b}\sum_{i=1}^{b}\log\frac{\exp\left(s\left(v_i,t_i\right)\right)}{\sum_{j=1}^{b}\exp\left(s\left(v_j,t_i\right)\right)}. \tag{4}$$

该损失从文本出发,鼓励将对应的视频排在所有视频中的首位.总损失 l 是文本到视频检索损失 l_{t2v} 和视频到文本检索损失 l_{v2t} 平均值,

$$l=\frac{1}{2}\left(l_{v2t}+l_{t2v}\right). \tag{5}$$

综合 2 个方向的损失,有效优化模型的双向检索能力.

3 实验与分析

3.1 数据集

实验在 2 个标准 VTR 数据集上进行测试.MSR-VTT[8] 包含来自 YouTube 的 10 000 个视频片段,涵盖广泛的场景和主题.采用“Training-7K”[14] 和“Training-9K”[11] 2 种数据集分割方式对模型进行测试.如无额外注释,消融实验部分采

用更为流行的“Training-9K”划分方式进行实验.LSMDC[20]由 118 081 个视频组成,这些视频从202部电影中提取,使用1 000个独立于训练集和验证集的视频进行测试.

3.2　评价指标

使用的评价指标如下:

1) R@K(recall at rank K)查询样本返回的前K个检索对象中找到正确样本的百分比,数值越大越好;

2) MdR(median rank)检索正确的样本在序列中位置的中位数,数值越小越好;

3) MnR(mean rank)检索正确的样本在序列中位置的平均数,数值越小越好.

3.3　实验设置

使用PyTorch库在单个NVIDIA A100 GPU上进行实验.所有实验都使用预训练的CLIP模型来初始化本文的视频帧编码器和文本编码器.除非另有说明,本文所有的模型都基于ViT-B/32骨干.Adam优化器[21]用于优化VTR模型,并使用余弦调度策略衰减学习率[22].微调过程中视频和文本编码器的初始学习率设置为1e-6,其他模块的学习率是3e-5.最大文本令牌长度为32,最大帧数为12,运行5个epoch,批量大小设置为32.联合跨模态交互模块的模态数量为2,隐藏层的维度为512,跨模态交互的层数为1,注意力头的数量为8.

3.4　实验结果与分析

为评估方法的有效性,在标准数据集MSR-VTT和LSMDC上进行测试.表1给出了本文方法与最先进模型在MSR-VTT数据集的“Training-7K”分割下的比较.表2是在“Training-9K”分割下的测试结果.表3反映模型在LSMDC数据集上的效果.

分析表1~表2的实验结果可知,本文基于联合跨模态注意力机制的方法在MSR-VTT数据集的不同划分方式上均能取得最优结果,表明本方法具有较强的适应性.结合LSMDC数据集上的实验结果,本文方法在多个数据集上均可以取得最优结果,证明其具有可扩展性.综合上述分析,充分证明本文方法的有效性.

3.5　消融实验

本节通过详细消融实验,以阐明主要参数设置对模型性能的影响.关于TopK文本-帧跨模态注意力机制的K选择,在MSR-VTT数据集上以“Training-9K”分割方式进行实验,采样的最大帧数是12.表4展示了不同K对模型检索准确率的影响.

表1　MSR-VTT-7K文本到视频检索结果
Table 1　Text-to-video retrieval results on MSR-VTT-7K

模型	R@1	R@5	R@10	MdR	MnR
ActBERT[16]	8.6	23.4	33.1	36.0	—
Howto100M[14]	14.9	40.2	52.8	9.0	—
ClipBERT[23]	22.0	46.8	59.9	6.0	—
All-in-one[24]	34.4	65.4	75.8	—	—
CLIP4Clip[2]	42.0	68.6	78.7	2.0	16.2
X-Pool[9]	43.9	72.5	82.3	2.0	14.6
Ours	44.6	73.1	84.0	2.0	12.4

表2　MSR-VTT-9K文本到视频检索结果
Table 2　Text-to-video retrieval results on MSR-VTT-9K

模型	R@1	R@5	R@10	MdR	MnR
MMT[11]	26.6	57.1	69.6	4.0	24.0
FROZEN[15]	32.5	61.5	71.2	3.0	—
All-in-one[24]	37.9	68.1	77.1	—	—
MAC[25]	38.9	63.1	73.9	3.0	—
Clover[26]	40.5	69.8	79.4	2.0	—
CLIP4Clip[2]	44.5	71.4	81.6	2.0	15.3
RVTR[27]	45.8	73.0	83.5	—	—
X-CLIP[3]	46.1	73.0	83.1	2.0	13.2
X-Pool[9]	46.9	72.8	82.2	2.0	14.3
STAN[7]	46.9	72.8	82.8	2.0	—
DGL[28]	47.0	70.4	81.0	—	16.4
TS2-Net[29]	47.0	74.5	83.8	2.0	13.0
TABLE[30]	47.1	74.3	82.9	2.0	13.4
Ours	47.2	73.1	84.3	2.0	11.3

表3　LSMDC文本到视频检索结果
Table 3　Text-to-video retrieval results on LSMDC

模型	R@1	R@5	R@10	MdR	MnR
MMT[11]	12.9	29.9	40.1	19.3	75.0
FROZEN[15]	15.0	30.8	39.8	20.0	—
RVTR[27]	19.2	38.0	47.0	—	—
DGL[28]	21.6	39.3	49.0	—	64.4
CLIP4Clip[2]	22.6	41.0	49.1	11.0	61.0
X-CLIP[3]	23.3	43.0	56.0	—	—
TS2-Net[29]	23.4	42.3	50.9	9.0	56.9
STAN[7]	23.7	42.7	51.8	9.0	—
TABLE[30]	24.3	44.9	53.7	8.0	52.7
Clover[26]	24.8	44.0	54.5	8.0	—
X-Pool[9]	25.2	43.7	53.5	8.0	53.2
Ours	25.4	43.8	54.2	8.0	52.8

表4 不同K的消融实验结果

Table 4 Ablation experiment results of different K

K	R@1	R@5	R@10	MdR	MnR
1	41.5	68.8	78.6	2.0	13.5
2	43.4	70.8	81.6	2.0	13.5
4	44.2	70.5	81.2	2.0	14.1
6	43.3	71.0	80.9	2.0	14.0
8	43.5	69.7	80.3	2.0	14.6
10	43.1	68.9	79.9	2.0	15.2
12	42.2	69.5	79.5	2.0	15.5

分析表4数据,以全部采样12次测试数据为基准,K取1时准确率明显下降,其原因是采样视频的1帧图像代表整个视频,无法涵盖视频中的丰富信息.K取值从2到10呈现起伏,但都要优于基线的测试结果,表明TopK文本-帧跨模态注意力机制是明显有效的.该类方法同样存在局限:K的选择会影响准确率,同时对于不同的数据集,最优的K也将会有所差异,这种不自适应的方法难以灵活应对多场景视频.影响最优K的最主要因素是数据集中视频内容的复杂度,复杂度高的视频最优K越大.

联合跨模态注意力机制模块用于建模相关性信息,同时对时间信息进行编码,跨模态编码器的层数同样影响实验效果.为此,表5为有无时间编码模块的联合跨模态注意力机制相较于基准的结果比较.

表5 时间信息编码实验结果

Table 5 Experimental results of temporal information encoding

类型	R@1	R@5	R@10	MdR	MnR
基准	44.5	71.4	81.6	2.0	15.3
无时间编码	46.8	72.5	82.4	2.0	14.2
有时间编码	47.2	73.1	84.3	2.0	11.3

分析表5可知,相较于基准方法,加入联合跨模态注意力机制能够显著提升模型的准确率,证明了相关性建模的有效性.加入时间编码模块的方法更具竞争力,证明模型对时间信息建模的有效性.

4 结 语

1)本文提出一种基于跨模态注意力机制的VTR方法,能够同时建模视频帧的时间信息和相关性信息,有效提升检索的准确率.

2)本文模型在多个标准数据集上取得具有竞争力的表现.

3)消融实验表明,本文提出的联合跨模态注意力机制中的时间建模模块和跨模态编码器对增强视频特征表示具有积极作用,能够为现实生活中的视频检索任务提供有力的理论支撑.

参考文献:

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]// Proceedings of the 38th International Conference on Machine Learning.Vienna: PMLR, 2021: 8748-8763.

[2] Luo H S, Ji L, Zhong M, et al. CLIP4Clip: an empirical study of CLIP for end to end video clip retrieval and captioning[J].*Neurocomputing*, 2022, 508: 293-304.

[3] Ma Y W, Xu G H, Sun X S, et al. X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval [C]// Proceedings of the 30th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2022: 638-647.

[4] Wu W H, Luo H P, Fang B, et al. Cap4Video: what can auxiliary captions do for text-video retrieval? [C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Vancouver BC: IEEE, 2023: 10704-10713.

[5] Fang H, Xiong P F, Xu L H, et al. Transferring image-CLIP to video-text retrieval via temporal relations[J].*IEEE Transactions on Multimedia*, 2023, 25: 7772-7785.

[6] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? [C]//Proceedings of the 38th International Conference on Machine Learning. Vienna: PMLR, 2021: 813-824.

[7] Liu R Y, Huang J J, Li G, et al. Revisiting temporal modeling for CLIP-based image-to-video knowledge transferring[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver BC: IEEE, 2023: 6555-6564.

[8] Xu J, Mei T, Yao T, et al. MSR-VTT: a large video description dataset for bridging video and language [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas: IEEE, 2016: 5288-5296.

[9] Gorti S K, Vouitsis N, Ma J W, et al. X-Pool: cross-modal language-video attention for text-video retrieval[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.New Orleans: IEEE, 2022: 5006-5015.

[10] Miech A, Laptev I, Sivic J. Learning a text-video embedding from incomplete and heterogeneous data [EB/OL]. (2018-04-07) [2024-10-24]. https://arxiv. org/pdf/1804.02516.pdf.

[11] Gabeur V, Sun C, Alahari K, et al. Multi-modal Transformer for video retrieval [C]// Computer Vision-ECCV 2020: 16th European Conference.Glasgow: Springer International Publishing, 2020: 214-229.

[12] Liu Y, Albanie S, Nagrani A, et al. Use what you have: video retrieval using representations from collaborative experts [EB/OL]. (2019-07-31) [2024-10-24]. https://arxiv.org/pdf/1907.13487.pdf.

[13] Jordan M I, Jacobs R A. Hierarchical mixtures of experts and the EM algorithm [J]. *Neural Computation*, 1994, 6 (2): 181-214.

[14] Miech A, Zhukov D, Alayrac J B, et al. HowTo100M: learning a text-video embedding by watching hundred million narrated video clips[C]// Proceedings of the IEEE/

CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 2630-2640.
[15] Bain M, Nagrani A, Varol G, et al. Frozen in time: a joint video and image encoder for end-to-end retrieval [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision.Montreal: IEEE, 2021: 1728-1738.
[16] Zhu L C, Yang Y. ActBERT: learning global-local video-text representations [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 8743-8752.
[17] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural Computation*, 1997,9(8): 1735-1780.
[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6000-6010.
[19] Buch S, Eyzaguirre C, Gaidon A, et al. Revisiting the "Video" in video-language understanding[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 2907-2917.
[20] Rohrbach A, Torabi A, Rohrbach M, et al. Movie description [J]. *International Journal of Computer Vision*, 2017, 123(1): 94-120.
[21] Kingma D P, Ba J. Adam: a method for stochastic optimization [EB/OL]. (2017-01-30) [2024-10-24]. https://arxiv.org/pdf/1412.6980.pdf.
[22] Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts[EB/OL].(2017-05-30)[2024-10-24]. https://arxiv.org/pdf/1608.03983.pdf.
[23] Zhao S, Zhu L C, Wang X H, et al. CenterCLIP: token clustering for efficient text-video retrieval[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid: ACM, 2022: 970-981.
[24] Wang J P, Ge Y X, Yan R, et al. All in one: exploring unified video-language pre-training[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver BC: IEEE, 2023: 6598-6608.
[25] Shu F X, Chen B L, Liao Y, et al. MAC: masked contrastive pre-training for efficient video-text retrieval[J]. *IEEE Transactions on Multimedia*, 2024, 26: 9962-9972.
[26] Huang J J, Li Y N, Feng J S, et al. Clover: towards a unified video-language alignment and fusion model [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver BC: IEEE, 2023: 14856-14866.
[27] Zhang H W, Yang Y, Qi F, et al. Robust video-text retrieval via noisy pair calibration[J]. *IEEE Transactions on Multimedia*, 2023, 25: 8632-8645.
[28] Yang X, Zhu L, Wang X, et al. DGL: dynamic global-local prompt tuning for text-video retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024: 6540-6548.
[29] Liu Y Q, Xiong P F, Xu L H, et al.TS2-net: token shift and selection transformer for text-video retrieval [C]// European Conference on Computer Vision. Cham: Springer, 2022: 319-335.
[30] Chen Y Z, Wang J, Lin L J, et al. Tagging before alignment: integrating multi-modal tags for video-text retrieval [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC: AAAI Press, 2023: 396-404.

(上接第 74 页)

[8] Deng C, Guo F H, Wen C Y, et al. Distributed resilient secondary control for DC microgrids against heterogeneous communication delays and DoS attacks [J]. *IEEE Transactions on Industrial Electronics*, 2022, 69 (11): 11560-11568.
[9] Chen Y L, Qi D L, Dong H N, et al. An FDI attack-resilient distributed secondary control strategy for islanded microgrids[J]. *IEEE Transactions on Smart Grid*, 2021, 12 (3): 1929-1938.
[10] Lian Z J, Guo F H, Wen C Y, et al. Distributed resilient optimal current sharing control for an islanded DC microgrid under DoS attacks [J]. *IEEE Transactions on Smart Grid*, 2021, 12(5): 4494-4505.
[11] Sahoo S, Mishra S, Peng J C, et al. A stealth cyber-attack detection strategy for DC microgrids[J]. *IEEE Transactions on Power Electronics*, 2019, 34(8): 8162-8174.
[12] Lu L Y, Liu H J, Zhu H, et al. Intrusion detection in distributed frequency control of isolated microgrids [J]. *IEEE Transactions on Smart Grid*, 2019, 10(6): 6502-6515.
[13] Jin X, Haddad W M, Yucelen T. An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems[J]. *IEEE Transactions on Automatic Control*, 2017, 62(11): 6058-6064.
[14] Zhou Q, Shahidehpour M, Alabdulwahab A, et al. A cyber-attack resilient distributed control strategy in islanded microgrids[J]. *IEEE Transactions on Smart Grid*, 2020, 11 (5): 3690-3701.
[15] Guo F H, Xu Q W, Wen C Y, et al. Distributed secondary control for power allocation and voltage restoration in islanded DC microgrids [J]. *IEEE Transactions on Sustainable Energy*, 2018, 9(4): 1857-1869.
[16] Zuo S, Altun T, Lewis F L, et al. Distributed resilient secondary control of DC microgrids against unbounded attacks [J]. *IEEE Transactions on Smart Grid*, 2020, 11 (5): 3850-3859.
[17] Nasirian V, Moayedi S, Davoudi A, et al. Distributed cooperative control of DC microgrids[J]. *IEEE Transactions on Power Electronics*, 2015, 30(4): 2288-2303.