文章编号:1672-3961(2025)01-0058-08 DOI:10.6040/j.issn.1672-3961.0.2023.172

基于跨模态注意力哈希学习的视频片段定位方法

谭智方[1],董飞[2],卢鹏宇[1],潘嘉男[1],聂秀山[1*],尹义龙[3]
(1.山东建筑大学计算机科学与技术学院, 山东 济南 250101; 2.山东师范大学新闻与传媒学院, 山东 济南 250014; 3.山东大学软件学院, 山东 济南 250100)

摘要:为提升视频片段定位的精度与检索效率,提出基于跨模态注意力哈希学习的视频片段定位方法。将查询语句和原始视频特征通过哈希学习模型转化成简洁的二值哈希码;使用软注意力模块对查询语句中的关键单词进行加权,将视频哈希码和查询语句哈希码输入一个增强的跨模态注意力模型中,挖掘视觉和语言之间的语义关系;设计一个得分预测和位置预测网络,对查询时刻的起始时间戳进行定位。在 2 个公开数据集上对所提方法进行试验验证,结果表明所提方法对检索效率提升约 7 倍。

关键词:视觉理解;视频片段定位;多模态检索;哈希学习;跨模态
中图分类号:TP37 **文献标志码**:A
引用格式:谭智方,董飞,卢鹏宇,等. 基于跨模态注意力哈希学习的视频片段定位方法[J]. 山东大学学报(工学版),2025,55(1):58-65.
TAN Zhifang, DONG Fei, LU Pengyu, et al. Video moment location method based on cross-modal attention hashing[J]. Journal of Shandong University (Engineering Science), 2025, 55(1):58-65.

Video moment location method based on cross-modal attention hashing

TAN Zhifang[1], DONG Fei[2], LU Pengyu[1], PAN Jianan[1], NIE Xiushan[1*], YIN Yilong[3]
(1. College of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, Shandong, China; 2. School of Journalism and Communication, Shandong Normal University, Jinan 250014, Shandong, China; 3. College of Software, Shandong University, Jinan 250100, Shandong, China)

Abstract: To enhance the accuracy of video segment localization and retrieval efficiency, a video moment location method based on cross-modal attention hashing was proposed. Query statements and original video features were transformed into concise binary hash codes through a hash learning model. A soft attention module was used to weight keywords in the query statements, and the video hash codes and query statement hash codes were input into an enhanced cross-modal attention model to explore semantic relationships between vision and language. A scoring prediction and position prediction network was designed to locate the starting timestamp at the query moment. Experimental validation of the proposed method on two public datasets showed that the proposed approach improved retrieval efficiency by approximately seven times.

Keywords: visual comprehension; video moment localization; multimodal retrieval; hashing; cross modal

0 引言

视频片段定位任务是在给定一个视频和一个查询语句的情况下,找到最符合查询语句的视频片段开始和结束时间戳。视频可以没有字幕,也可以没有声音,查询语句可以是非结构化的自然语言。视频片段定位是一项非常具有挑战性的任务:一方面,要求计算机理解 2 种模态视频和文本的语义,建立匹配关系;另一方面,处理视频往往需要大量计算,对检索效率产生负面影响。

现有工作大多聚焦于对视频片段检索定位精度的研究上,主要利用模态之间的细粒度匹配关系提高准确性,例如:文献[1]提出时序上下文关联网

收稿日期:2023-07-17
基金项目:国家自然科学基金资助项目(62176141, 62102235);山东省泰山学者资助项目(tsqn202103088);山东省自然科学基金资助项目(ZR2020QF029)
第一作者简介:谭智方(1997—),男,山东潍坊人,硕士研究生,主要研究方向为计算机视觉中的图像处理。E-mail:826133130@qq.com
*通信作者简介:聂秀山(1981—),男,江苏徐州人,教授,博士生导师,博士,主要研究方向为计算机视觉。E-mail:niexiushan@163.com

络,通过计算查询语句与视频不同尺度的不同部分之间的相似度粗略定位片段,需要大量计算,不够精确;文献[2]提出一种基于滑动窗口的模型,可以微调窗口内的边界,实现更细粒度的定位。为了进一步提高准确性,跨模态注意机制被广泛使用[3]。文献[4]建议在早期阶段将文本与视频 2 种模态相互作用,以产生语义上更丰富的建议;文献[5]提出一种同时提议和推理的网络。为了进一步减少与候选或滑动窗口相关的计算工作量,文献[6]使用深度强化学习自动找到最佳窗口;文献[7]设计了一种用于感知边界的方法。

随着数据量增加,基于自然语言查询语句的视频片段定位任务对检索效率的要求也越来越高,因此,如何在保证定位精度的同时加快检索速度至关重要。为解决这一问题,本研究把哈希学习引入视频片段定位领域。哈希学习是通过机器学习模型把高维的数据特征转变为简洁的二值哈希码表示,可以有效提升视频检索速度,减少内存存储。基于自然语言查询的视频片段定位任务本质上属于跨模态检索领域,通过哈希学习可以把视频和文本映射到同一个海明空间(海明空间即各数据之间以海明距离区别的空间,海明距离是指 2 个字符串对应位置不同字符的个数),利用不同模态数据交互和分析,因此,本研究基于哈希学习算法完成视频片段定位任务。

1　相关工作

1.1　视频动作定位

视频动作定位是定位视频中某些动作开始和结束时间的过程。与本研究任务相比,视频动作定位不能使用自然语言进行查询,动作类别有限,但成熟的模型经常作为其他视频任务的骨干网络。早期的工作通过手动聚合执行框架或窗口级别的分类进行定位[8-9],后来使用提案生成和边界微调的两阶段方法[10-11]。一些模型通常将提案生成和边界微调结合起来进行端到端的训练[12-13]。视频动作定位的一阶段方法是直接在视频帧中进行动作检测和定位。

1.2　基于文本查询的视频检索

基于文本查询的视频检索是从与文本描述相关联的视频集合中找到整个视频。与本研究中的片段定位不同,基于文本查询的视频检索不需要预测时刻的开始和结束时间戳,因此,主要困难是学会区分不同的视频,而不是同一视频的不同部分。目前,这一任务的主要方法是将不同的模态特征编码到联合嵌入空间中,以测量语义相似度。文献[14]将视频和文本编码成全局向量,尽管这种全局表示是有效的,但可能导致一些关键细节丢失。为避免这些问题,文献[15]考虑计算单词和帧之间的匹配关系,进一步计算整个视频和查询之间的匹配关系。然而,考虑自然语言通常包含复杂逻辑结构,有时部分匹配并不代表整体匹配关系,文献[16]中的一些方法一定程度上解决了此类问题。

1.3　哈希学习方法

近年来,哈希学习方法作为一类高效的近似近邻方法,广泛用于大规模图像检索中。哈希学习方法可以将异构的高维数据压缩成紧凑的二进制码,同时保持原始样本空间的相似性。现有的哈希学习方法一般可以分为 2 类:数据独立的方法和数据依赖的方法。对于数据独立的方法,哈希函数独立于训练数据,代表性方法是局部敏感哈希法(locality-sensitive hashing, LSH)[17]和相应的变体[18-20]。数据依赖的方法也称为学习哈希(learning to hash, L2H)方法,从训练数据中学习哈希函数。与数据独立的方法相比,数据依赖的方法通常可以从数据中学习内在属性,使用紧凑的二进制代码获得更好的性能。根据标签利用率,L2H 方法可以分为 2 种类型:无监督哈希方法[21-24]和有监督哈希方法[25-29]。谱哈希(spectral hashing, SH)[30-31]、基于主成分分析的哈希(principal component analysis hashing, PCAH)[32]和迭代量化(iterative quantization, ITQ)[33]是经典无监督哈希方法。无监督哈希方法不利用训练数据的标签信息,可能导致模式分类中的信息丢失。因此,已经提出许多有监督哈希方法,例如监督离散哈希法(supervised discrete hashing, SDH)[26-27]、快速监督离散哈希法(fast supervised discrete hashing, FSDH)[34]、松弛监督离散哈希法(supervised discrete hashing with relaxation, SDHR)[35]和基于语义标签回归的快速离散跨模态哈希法[28],充分利用类标签作为监督学习哈希码。一般相比无监督哈希方法,有监督哈希方法获得的检索性能更高。

2　本研究方法

2.1　问题描述

给定一个原始视频和查询语句,视频片段定位任务的目标是从视频中找到时刻的开始时间戳 t_s 和结束时间戳 t_e。时刻对应查询语句的语意。一个视频

可以看成由 T 个连续片段组成,在使用深度神经网络提取语意特征后可以表示为特征序列 $\boldsymbol{V}=\{\boldsymbol{c}_t\}_{t=1}^{T}$,其中 $\boldsymbol{c}_t$ 为第 t 时间步下的特征。文本信息使用 Glove 提取特征后[36],N 个词的语义特征可以表示为 $\boldsymbol{S}=\{\boldsymbol{w}_n\}_{n=1}^{N}$,其中 $\boldsymbol{w}_n$ 为第 n 个单词的词向量。

2.2 模型结构

本研究主要思路是构建一个模型,可以为视频 v 生成一组哈希码 $\boldsymbol{H}_v=\{\boldsymbol{h}_t^v\}_{t=1}^{T}$,为查询语句 s 生成一串哈希码 $\boldsymbol{h}^s$,其中 $\boldsymbol{h}_t^v$ 为视频时间步长 t 的哈希码。通过跨模态注意力模块对关键部分信息进行加权,根据得分预测和位置定位网络得到预测的起始时间戳。本研究方法的框架如图 1 所示,共包括 4 个组成部分:视频哈希网络、文本哈希网络、跨模态注意力模块及位置定位网络。

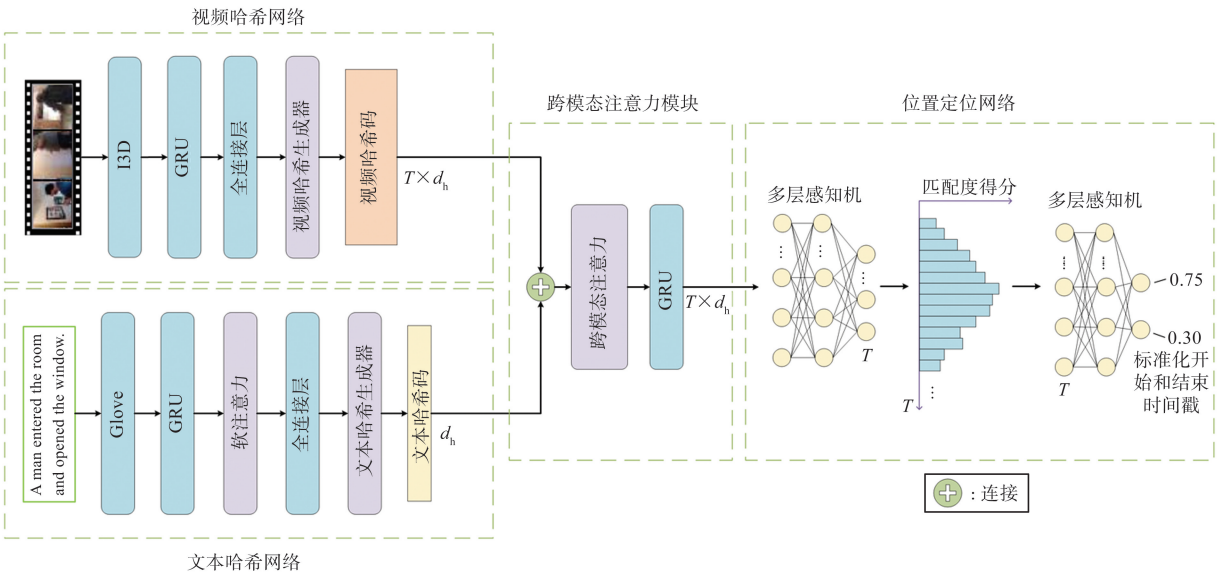

图 1 总体模型结构示意图

Fig.1 The overall model structure diagram

2.2.1 视频哈希网络

对于原始视频输入,使用视频编码器 I3D 将其转换为特征序列 $\boldsymbol{V}=\{\boldsymbol{c}_t\}_{t=1}^{T}$[37]。I3D 是一种基于 3D 卷积的特征提取网络,本研究借助 I3D 提取视频特征。将这些特征通过双向门控循环单元(gated recurrent unit, GRU)网络挖掘视频的时序信息[38],使用具有激活函数的全连接层(fully connected layers, FC)生成每个时刻的向量,离散化的向量是视频的哈希码。生成哈希码的过程如下:

$$\boldsymbol{h}_t^{g1}=\mathrm{GRU}(\boldsymbol{c}_t,\boldsymbol{h}_{t-1}^{g1}),\tag{1}$$

$$\boldsymbol{r}_t^v=\mathrm{Tanh}(\boldsymbol{W}_\alpha\boldsymbol{h}_t^{g1}+\boldsymbol{b}_\alpha),\tag{2}$$

$$\boldsymbol{h}_t^v=\mathrm{sign}(\boldsymbol{r}_t^v),\tag{3}$$

式中,$\boldsymbol{h}_t^{g1}$ 为 GRU 中时间步长 t 时双向输出拼接后的向量,$\boldsymbol{r}_t^v$ 为实数的归一化向量,sign(·)表示将连续的值映射为离散的值,$\boldsymbol{W}_\alpha$ 和 $\boldsymbol{b}_\alpha$ 分别为可学习的权重矩阵和偏差项,Tanh(·)为非线性激活函数。

2.2.2 文本哈希网络

文本哈希网络的结构与视频哈希网络相似,使用 Glove 提取特征,对文本特征进行处理。但考虑查询语句中词语的重要度不同,本研究使用软注意力模块获得一个增强的句子特征 $\boldsymbol{s}_{ph}$,通过使用软注意力模块对 GRU 最后时间步的输出进行加权。文本特征处理过程如下:

$$\boldsymbol{h}_n^{g2}=\mathrm{GRU}(\boldsymbol{w}_t,\boldsymbol{h}_{n-1}^{g2}),\tag{4}$$

$$\boldsymbol{\mu}_n=\boldsymbol{W}_\mu\boldsymbol{h}_n^{g2}+\boldsymbol{b}_\mu,\tag{5}$$

$$a_n=\frac{\exp(\boldsymbol{\mu}_n)}{\sum_{n=1}^{N}\exp(\boldsymbol{\mu}_n)},\tag{6}$$

$$\boldsymbol{s}_{ph}=\sum_{n=1}^{N}a_n\boldsymbol{h}_n^{g2},\tag{7}$$

式中,$\boldsymbol{h}_n^{g2}$ 为 GRU 中第 n 步的隐藏状态,$\boldsymbol{W}_\mu$ 和 $\boldsymbol{b}_\mu$ 为可学习的权重矩阵和偏差项,$\boldsymbol{\mu}_n$ 为线性变换后的隐藏状态,a_n 为获得的注意力权重。增强句子特征输出的实数向量

$$\boldsymbol{r}^s=\mathrm{Tanh}(\boldsymbol{W}_\beta\boldsymbol{s}_{ph}+\boldsymbol{b}_\beta),\tag{8}$$

式中,$\boldsymbol{W}_\beta$ 和 $\boldsymbol{b}_\beta$ 分别为可学习的权重矩阵和偏差项。查询语句的哈希码

$$\boldsymbol{h}^s=\mathrm{sign}(\boldsymbol{r}^s)。\tag{9}$$

2.2.3 跨模态注意力模块

为了关注跨模态间的关系及时序信息,本研究提出跨模态注意力模块,可以根据句子特征调整视频片段特征的权重。

为便于计算视频片段和查询语句间的相似性,本研究将视频特征 $\boldsymbol{h}_t^v$ 和查询语句特征 $\boldsymbol{h}^s$ 映射到相同维度,具有相同维度的视频语句特征 $\boldsymbol{q}_t$ 和查询语句的特征 $\boldsymbol{k}_t$ 分别为:

$$\boldsymbol{q}_t=\boldsymbol{W}_q\boldsymbol{h}_t^v, \tag{10}$$

$$\boldsymbol{k}_t=\boldsymbol{W}_k\boldsymbol{h}^s, \tag{11}$$

式中,$\boldsymbol{W}_q$ 和 $\boldsymbol{W}_k$ 为映射矩阵。降维后,新的视频片段特征

$$\boldsymbol{v}_t=\boldsymbol{W}_v\boldsymbol{h}_t^v+\boldsymbol{b}_v, \tag{12}$$

式中,$\boldsymbol{W}_v$ 和 $\boldsymbol{b}_v$ 分别为可学习的权重矩阵和偏差项。

通过查询语句对 $\boldsymbol{v}_t$ 进行加权,得到加权后的视频片段特征

$$\boldsymbol{c}_{\mathrm{hs}}^t=\boldsymbol{v}_t\mathrm{e}^{r_t}, \tag{13}$$

式中,r_t 为视频和文本间对应程度的反馈,$r_t=\mathrm{Tanh}(\boldsymbol{q}_t\cdot\boldsymbol{k}_t)$,会根据视频片段和查询语句之间的相关性进行调整,其值在-1 和 1 之间变化。将加权后的特征送入 GRU 网络中。这里 GRU 网络的作用是记忆特征中的重要信息,遗忘不重要的特征信息。最终的输出将输入至最终的位置定位网络。跨模态注意力模块最终输出结果表示为 $\boldsymbol{H}_m=\{\boldsymbol{h}_m^t\}_{t=1}^T$,其中 $\boldsymbol{h}_m^t$ 为第 t 个时间步长下的视频注意力特征,m 为维度。

2.2.4　位置定位网络

在位置定位网络中,本研究设计一个分数预测器,是一个带有 Sigmoid 和 Relu 激活函数的多层感知机(multilayer perceptron, MLP),输出为每一个时间步的匹配得分,得分是 0~1 中的一个实数。位置定位器是一个带有 Tanh 激活函数的 MLP,接收所有时间步的分数,输出查询时刻的位置索引。

将跨模态注意力模块的输出结果 $\boldsymbol{H}_m=\{\boldsymbol{h}_m^t\}_{t=1}^T$ 作为输入,在每个时间步计算匹配度分数

$$s_p^t=\mathrm{Sigmoid}(\boldsymbol{h}_M^t), \tag{14}$$

式中:Sigmoid(·)为激活函数;$\boldsymbol{h}_M^t$ 为归一化的视频注意力特征 $\boldsymbol{h}_{m'}^t$ 在使用激活函数增加非线性后的结果,$\boldsymbol{h}_M^t=\boldsymbol{W}_\theta(\mathrm{Relu}(\boldsymbol{W}_\gamma\boldsymbol{h}_{m'}^t+\boldsymbol{b}_\gamma))+\boldsymbol{b}_\theta$,$\boldsymbol{h}_{m'}^t=\mathrm{Relu}(\mathrm{Batchnorm}(\boldsymbol{h}_m^t))$,其中 $\boldsymbol{W}_\theta$、$\boldsymbol{W}_\gamma$、$\boldsymbol{b}_\theta$、$\boldsymbol{b}_\gamma$ 为可学习的权重矩阵和偏差项,Relu(·)为激活函数,Batchnorm(·)为批量归一化函数。所有时间步的 s_p^t 集合为一个得分向量 $\boldsymbol{s}_T$。

将 $\boldsymbol{s}_T$ 输入至位置定位器,即最后一个多层感知机。查询到时刻的位置索引集合

$$\boldsymbol{L}_{\mathrm{p}}=\boldsymbol{W}_\delta(\mathrm{Tanh}(\boldsymbol{W}_\varepsilon\boldsymbol{s}_T+\boldsymbol{b}_\varepsilon))+\boldsymbol{b}_\delta, \tag{15}$$

式中,$\boldsymbol{W}_\delta$、$\boldsymbol{W}_\varepsilon$ 为可学习的权重矩阵,$\boldsymbol{b}_\varepsilon$、$\boldsymbol{b}_\delta$ 为偏差项。$\boldsymbol{L}_{\mathrm{p}}$ 中第一项为开始片段的索引,第二项为结束片段的索引。

2.3　训练和推理

2.3.1　训练

给定一个视频和一个查询语句,相应的开始时间戳和结束时间戳分别为 t_{s} 和 t_{e}。如果一个片段对应查询语句所描述的事件,则其相关性得分设置为 1,否则设置为 0。

本研究采用交叉熵损失作为得分损失

$$L_{\mathrm{sc}}=-\frac{1}{T}\sum_{t=1}^{T}s_{\mathrm{gt}}^t\ln(s_{\mathrm{p}}^t)+(1-s_{\mathrm{gt}}^t)\ln(1-s_{\mathrm{p}}^t), \tag{16}$$

式中 s_{gt}^t 为真实值对应的每个片段的得分。

位置损失

$$L_{\mathrm{loc}}=\frac{1}{2}(l^{\mathrm{s}}-t_{\mathrm{s}}^n)^2+\frac{1}{2}(l^{\mathrm{e}}-t_{\mathrm{e}}^n)^2, \tag{17}$$

式中,t_{s}^n 和 t_{e}^n 分别为标准化的开始时间戳和结束时间戳(真实值),l^{s} 和 l^{e} 分别为预测的开始索引和结束索引。

总损失

$$L=L_{\mathrm{sc}}+\lambda L_{\mathrm{loc}}, \tag{18}$$

式中 λ 为调节损失权重的超参数。

2.3.2　推理

推理过程可以端到端完成,或通过相应网络生成用于存储的哈希码,在需要时使用位置预测网络进行匹配。在获得 l^{s} 和 l^{e} 后,预测开始时间戳 t_{s}^* 和预测结束时间戳 t_{e}^* 可以通过以下计算获得:

$$t_{\mathrm{s}}^*=l^{\mathrm{s}}d, \tag{19}$$

$$t_{\mathrm{e}}^*=l^{\mathrm{e}}d, \tag{20}$$

式中 d 为视频时长。

3　试验

3.1　数据集

本研究在 2 个公开的基准数据集上评估所提模型,即 Charades-STA 数据集及 ActivityNet Captions 数据集。

3.1.1　Charades-STA 数据集

Charades-STA 数据集用半自动方法进行标注[2],因此句子通常较短,在形式上更简单,包含 13 898个匹配的查询对,每个视频的平均长度为 31 s,每个时刻大约持续 8 s。

3.1.2　ActivityNet Captions 数据集

ActivityNet Captions 数据集中共 20 000 个视频,平均每个视频包含 3.65 个句子,每个句子都对

应视频的某个时刻,时长可以从几秒到一百多秒[39],句子本身也非常复杂,包含多个连续动作。

3.2 评价指标

本研究使用预测时间和真实时间之间交集和并集之比 I_{oU} 作为评价指标。为了公平比较,本研究以"R@1, $I_{oU}=y$"作为评估标准,认定至少有1个预测结果与真实结果的 $I_{oU}>y$ 时为预测正确,"R@1, $I_{oU}=y$"的结果为预测正确的结果在测试集中的占比。

3.3 试验设置

除特别说明,本试验中2个数据集超参数的设置相同。试验设置在 Ubuntu 16.04 上的一个 Nvidia 2080Ti GPU 上运行,内存为 512 GB。视频和句子哈希码均为64位。在 ActivityNet Captions 数据集和 Charades-STA 数据集上分别使用500维的C3D特征[40]和1 024维的I3D特征[37]作为视频特征。查询语句采用300维的 Golve 特征。GRU 和 FC 的隐藏层大小分别为256和128。此外,λ 设置为0.01。本试验把 Charades-STA 中的视频平均采样为64个片段,把 ActivityNet Captions 数据集中的视频采样为128个片段。所有试验均使用 Adam 优化器进行,学习率为0.001,批次大小为64。

3.4 模型精度

在2个数据集上,本研究提出的模型和其他模型之间的比较如表1、2所示,其中最优结果加粗表示。在2个数据集上还提供了使用真实特征的试验结果。试验结果表明,本研究模型的定位精度高于现有模型的精度。

表1 ActivityNet Captions 数据集的 R@1 性能比较

Table 1 R@1 performance comparison for the ActivityNet Captions dataset 单位:%

模型	R@1		
	$I_{oU}=0.3$	$I_{oU}=0.5$	$I_{oU}=0.7$
MCN[1]	39.35	21.36	6.43
CTRL[2]	47.43	29.01	10.34
TGN[3]	45.51	28.47	
TripNet[19]	48.42	32.19	13.93
ACRN[20]	49.70	31.67	11.25
VMLH[41]	52.15	**34.50**	**17.16**
本研究模型	**52.21**	34.37	17.11

注:时刻上下文网络(moment context network, MCN)、跨模态时序回归模型(cross-modal temporal regression localizer, CTRL)、时序定位网络(temporal ground net, TGN)、三重损失网络(triple loss network, TripNet)、跨模态注意力检索网络(attentive cross-modal retrieval network, ACRN)、哈希视频时刻定位(video moment location via hashing, VMLH)。

表2 Charades-STA 数据集的 R@1 性能比较

Table 2 R@1 performance comparison for the Charades-STA dataset 单位:%

模型	R@1	
	$I_{oU}=0.5$	$I_{oU}=0.7$
CTRL[2]	23.63	8.89
MLVI[4]	35.60	15.80
ACL-K[42]	30.48	12.20
ACRN[20]	20.26	7.64
SM-RL[6]	24.36	11.17
QSPN[4]	35.60	15.80
TripNet[19]	36.61	14.50
VMLH[41]	43.80	20.32
本研究模型	**43.97**	**21.07**

注:视觉和语言特性的多级模型(multilevel language and vision integration, MLVI)、基于动作概念的定位器模型(activity concepts based localizer, ACL-K)、语义匹配的强化学习模型(semantic matching reinforcement learning, SM-RL)、查询导向的提案网络(query-guided segment proposal network, QSPN)。

3.5 模型效率

本研究方法把哈希学习引入视频片段定位任务中,在保证定位精度的同时,可以有效提升检索速度。为验证这一结论,本试验测试了单次查询平均时间,试验结果如表3所示。表3展示了不同设置和相同批次下模型的效率,其中批次都为1。由表3可以看出,因使用哈希码,模型的计算量变为原来的1/7。如果考虑大批量,每次检索所需的平均时间可以减少到微秒级。

表3 模型间单次查询平均时间比较

Table 3 Comparison of average query time between models 单位:s

模型	单次查询平均时间
CTRL[2]	3.410 0
ACRN[20]	4.420 0
ABLR[43]	0.060 0
CMHN[44]	0.007 6
Ours-full	0.011 7
Ours-vh	0.006 4
Ours-h	0.002 7

注:协同注意力回归模型(attention based location regression, ABLR)、跨模态哈希网络(cross-modal hashing network, CMHN)。Ours-full 表示输入视频和句子都不预先存储为哈希码;Ours-vh 表示视频预存为哈希码,句子需要经过文本哈希网络;Ours-h 表示都使用预存的哈希码进行检索。

3.6 消融试验

为评估哈希生成器对效率提升的有效性,在2个数据集上进行消融试验,结果如表4所示。由表4可知,相比于哈希生成器生成32、64、128位哈希码,没有哈希生成器时效率明显下降。与本研究采

用 64 位哈希码相比,使用哈希生成器的效率明显提升了近 7 倍。另外,本研究进行了不同长度哈希码的定位精度对比试验,结果如表 5 所示。由表 5 可知:使用 32 位哈希码时,定位精度明显下降;相比 128 位,64 位的定位精度略有下降,但效率更优。所以,本研究在精度下降的可接受范围内采取 64 位哈希码作为哈希生成器设置。哈希码带来的特征损失无法消除,所以在精度上也有相应表现,哈希码的表征性不强也是本研究的局限性之一。该消融试验证明哈希生成器显著提升了查询效率。

表 4　哈希定位效率消融试验比较

Table 4　Comparison of hash localization efficiency ablation experiments　　单位:s

哈希码长度	Charades-STA 数据集平均定位时间	ActivityNet Captions 数据集平均定位时间
32 位	0.008 9	0.009 2
64 位	0.009 7	0.013 6
128 位	0.018 6	0.023 2
None	0.072 6	0.087 5

注:None 表示不采用哈希生成器。

表 5　哈希定位精度消融试验比较

Table 5　Comparison of hash localization accuracy ablation experiments　　单位:%

哈希码长度	Charades-STA 数据集定位精度		ActivityNet Captions 数据集定位精度		
	$I_{oU}=0.5$	$I_{oU}=0.7$	$I_{oU}=0.3$	$I_{oU}=0.5$	$I_{oU}=0.7$
32 位	41.57	17.32	50.11	30.23	14.80
64 位	43.97	21.07	52.21	34.37	17.11
128 位	44.03	21.46	52.39	34.39	17.28
None	44.59	22.26	52.69	34.98	18.18

4　结论

本研究提出一种基于哈希学习的高效视频片段定位网络,使用跨模态的哈希方法解决视频片段定位问题,在提高定位精度的同时,可以加快检索速度并减少存储空间,为大规模的视频片段定位提供可行的思路。在未来工作中,将对哈希码在视频和文本语意更清晰的表征上做进一步研究。

参考文献:

[1] HENDRICK L A, WANG O, SHECHTMAN E, et al. Localizing moments in video with natural language[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 5804-5813.

[2] GAO J, SUN C, YANG Z, et al. TALL: temporal activity localization via language query[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 5277-5285.

[3] CHEN J, CHEN X, MA L, et al. Temporally grounding natural sentence in video[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2018: 162-171.

[4] XU H, HE K, PLUMMER B A, et al. Multilevel language and vision integration for text-to-clip retrieval [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI, 2019: 9062-9069.

[5] ZHANG D, DAI X, WANG X, et al. MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 1247-1257.

[6] WANG W, HUANG Y, WANG L. Language-driven temporal activity localization: a semantic matching reinforcement learning model[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 334-343.

[7] GHOSH S, AGARWAL A, PAREKH Z, et al. ExCL: extractive clip localization using natural language descriptions[EB/OL]. (2019-04-04)[2023-11-12]. https://arxiv.org/pdf/1904.02755.

[8] SHOU Z, CHAN J, ZAREIAN A, et al. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 1417-1426.

[9] ZENG R, GAN C, CHEN P, et al. Breaking winner-takes-all: iterative-winners-out networks for weakly supervised temporal action localization[J]. IEEE Transactions on Image Processing, 2019, 28(12): 5797-5808.

[10] SHOU Z, WANG D, CHANG S F. Temporal action localization in untrimmed videos via multi-stage CNNs [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 1049-1058.

[11] XU H, DAS A, SAENKO K. R-C3D: region convolutional 3D network for temporal activity detection [C]//Proceedings of the 2017 IEEE International

Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 5794-5803.

[12] LIN T, ZHAO X, SHOU Z. Single shot temporal action detection[C]// Proceedings of the 25th ACM International Conference on Multimedia. New York, USA: ACM, 2017: 988-996.

[13] WANG J, CHENG Y, FERIS R S. Walk and learn: facial attribute representation learning from egocentric video and contextual data[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 2295-2304.

[14] MITHUN N C, LI J, METZE F, et al. Learning joint embedding with multimodal cues for cross-modal video-text retrieval[C]//Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. New York, USA: ACM, 2018: 19-27.

[15] YU Y, KIM J, KIM G. A joint sequence fusion model for video question answering and retrieval[C]// Proceedings of the European Conference on Computer Vision (ECCV). Piscataway, USA: IEEE, 2018: 471-487.

[16] CHEN S, ZHAO Y, JIN Q, et al. Fine-grained video-text retrieval with hierarchical graph reasoning[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 10635-10644.

[17] GIONIS A, INDYK P, MOTWANI R. Similarity search in high dimensions viahashing[C]//Proceedings of the International Conference on Very Large Data Bases. New York, USA: Springer, 1999: 518-529.

[18] DATAR M, IMMORLICA N, INDYK P, et al. Locality-sensitivehashing scheme based on p-stable distributions[C]//Proceedings of the Twentieth Annual Symposium on Computational Geometry. New York, USA: ACM, 2004: 253-262.

[19] ANDONI A, INDYK P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions[C]//Proceedings of the 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS′06). Berkeley, USA: IEEE, 2006: 459-468.

[20] KULIS B, GRAUMAN K. Kernelized locality-sensitive hashing for scalable image search[C]//Proceedings of the 2009 IEEE 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009: 2130-2137.

[21] LUO W, LIU W, GAO S. A revisit of sparse coding based anomaly detection in stacked RNN framework [C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 341-349.

[22] LIU L, SHAO L. Sequential compact code learning for unsupervised image hashing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(12): 2526-2536.

[23] ZHU L, SHEN J, XIE L, et al. Unsupervised visual hashing with semantic assistant for content-based image retrieval[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(2): 472-486.

[24] ZHU L, HUANG Z, LI Z, et al. Exploring auxiliary context: discrete semantic transfer hashing for scalable image retrieval[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(11): 5264-5276.

[25] LI G, SHEN C, VAN DEN HENGEL A. Supervised hashing using graph cuts and boosted decision trees[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(11): 2317-2331.

[26] WANG Q, ZHANG Z, SI L.Ranking preserving hashing for fast similarity search[C]//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI, 2015: 3911-3917.

[27] SHEN F, SHEN C, LIU W, et al. Supervised discrete hashing[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015: 37-45.

[28] LIU X, NIE X, ZENG W, et al. Fast discrete cross-modal hashing with regressing from semantic labels [C]//Proceedings of the 26th ACM International Conference on Multimedia. New York, USA: ACM, 2018: 1662-1669.

[29] GUI J, LI P. R2SDH:robust rotated supervised discrete hashing[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, USA: ACM, 2018: 1485-1493.

[30] WEISS Y, TORRALBA A, FERGUS R. Spectral hashing[C]// Proceedings of the 21st International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2008: 1753-1760.

[31] LIU Q, LIU G, LI L, et al. Reversed spectral hashing [J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 29(6): 2441-2449.

[32] HU Z, PAN G, WANG Y, et al. Sparse principal component analysis via rotation and truncation[J]. IEEE Transactions on Neural Networks and Learning Systems,

2015, 27(4): 875-890.
[33] GONG Y, LAZEBNIK S, GORDO A, et al. Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(12): 2916-2929.
[34] GUI J, LIU T, SUN Z, et al. Fast supervised discrete hashing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(2): 490-496.
[35] GUI J, LIU T, SUN Z, et al. Supervised discrete hashing with relaxation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 29(3): 608-617.
[36] PENNINGTON J, SOCHER R, MANNING C D. GloVe: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, USA: ACL, 2014: 1532-1543.
[37] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017: 6299-6308.
[38] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11)[2023-11-12]. https://arxiv.org/pdf/1412.3555.
[39] KRISHNA R, HATA K, REN F, et al. Dense-captioning events in videos[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 706-715.
[40] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015: 4489-4497.
[41] TAN Z, DONG F, LIU X, et al. VMLH: efficient video moment location via hashing[J]. Electronics, 2023, 12(2): 420.
[42] GE R, GAO J, CHEN K, NEVATIA R. MAC: mining activity concepts for language-based temporal localization[C]//Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2019: 245-253.
[43] YUAN Y, MEI T, ZHU W. To find where you talk: temporal sentence localization in video with attention based location regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, USA: AAAI, 2019: 9159-9166.
[44] HU Y, LIU M, SU X, et al. Video moment localization via deep cross-modal hashing[J]. IEEE Transactions on Image Processing, 2021, 30: 4667-4677.

(编辑:孙亚彤)

(上接第 57 页)

[16] HIREMATH P S, SHIVASHANKAR S. Wavelet based co-occurrence histogram features for texture classification with an application to script identification in a document image[J]. Pattern Recognition Letters, 2008, 29(9): 1182-1189.
[17] HAN X, AYSA A, MAMAT H, et al. Script identification of central Asia based on fused texture features[C]//Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR). Beijing, China: IEEE, 2018: 3675-3680.
[18] RAJPUT G G, UMMAPURE S B. Script identification from handwritten document images using LBP technique at block level[C]//Proceedings of the 2019 International Conference on Data Science and Communication (IconDSC). Bangalore, India: IEEE, 2019: 8816944.
[19] HARALICK R M, SHANMUNGAM K, DINSTEIN I. Textural features of image classification[J]. IEEE Transactions on Systems, Man and Cybernatics, 1973, 3: 610-621.
[20] SINGH P K, DALAL S K, SARKAR R, et al. Page-level script identification from multi-script handwritten documents[C]//Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT). Hooghly, India: IEEE, 2015: 7060113.
[21] NAGHASHI V. Co-occurrence of adjacent sparse local ternary patterns: a feature descriptor for texture and face image retrieval[J]. Optik, 2018, 157: 877-889.
[22] TIAN S, BHATTACHARYA U, LU S, et al. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients[J]. Pattern Recognition, 2016, 51: 125-134.
[23] NANNI L, BRAHNAM S, LUMINI A. Selecting the best performing rotation invariant patterns in localbinary/ternary patterns[C]// Proceedings of the 2010 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV′10). Las Vegas, USA: IEEE, 2010: 369-375.
[24] LI S, MUTELIPU M, MAMAT H, et al. Script identification of multi-script document images based on discrete curvelet transform[J]. Computer Engineering and Design, 2019, 40(5): 1376-1382.

(编辑:孙亚彤)