软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(7):1969–1979 [doi: 10.13328/j.cnki.jos.005935]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于全局和局部信息的视频记忆度预测*

王 帅, 王维莹, 陈师哲, 金 琴

(中国人民大学 信息学院,北京 100872)

通讯作者: 金琴, E-mail: qjin@ruc.edu.cn

摘 要: 视频的记忆度是一种度量指标,用来表示一段视频能够普遍被人记住的程度.令人记忆深刻而难忘的视频具有很大的潜在价值,因此对能够进行大规模视频记忆度自动预测的模型将会有广大的应用前景和市场,例如视频检索、数字内容推荐、广告设计、教育系统等等.现有的大部分工作都是直接利用深度神经网络学习到的一个全局表示来进行记忆度的预测,没有给予局部细节足够的重视.提出了一个基于全局和局部信息的视频记忆度预测模型,其中,包含 3 个模块:全局性的上下文表示模块、空间布局表示模块和局部的物体注意力模块.在实验结果中,全局性的上下文表示模块和局部的物体注意力模块分别具有很好的表现.而空间布局表示模块的预测能力虽不如其他两个模块,但 3 个模块的融合使结果有了进一步的提升.最后,在 MediaEval 2018 Media Memorability Prediction Task 的数据集上证明了模型的有效性.

关键词: 视频记忆度;注意力机制;物体检测;神经网络

中图法分类号: TP391

中文引用格式: 王帅,王维莹,陈师哲,金琴.基于全局和局部信息的视频记忆度预测.软件学报,2020,31(7):1969–1979. http://www.jos.org.cn/1000-9825/5935.htm

英文引用格式: Wang S, Wang WY, Chen SZ, Jin Q. Video memorability prediction based on global and local information. Ruan Jian Xue Bao/Journal of Software, 2020,31(7):1969–1979 (in Chinese). http://www.jos.org.cn/1000-9825/5935.htm

Video Memorability Prediction Based on Global and Local Information

WANG Shuai, WANG Wei-Ying, CHEN Shi-Zhe, JIN Qin

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: Memorability of a video is a metric to describe that how memorable the video is. Memorable videos contain huge values and automatically predicting the memorability of large numbers of videos can be applied in various applications including digital content recommendation, advertisement design, education system, and so on. This study proposes a global and local information based framework to predict video memorability. The framework consists of three components, namely global context representation, spatial layout, and local object attention. The experimental results of the global context representation and local object attention are remarkable, and the spatial layout also contributes a lot to the prediction. Finally, the proposedmodel improves the performances of thebaseline of MediaEval 2018 Media Memorability Prediction Task.

Key words: video memorability, attention, object detection, neural network

随着互联网、移动设备、存储技术等技术的发展,多媒体内容越来越成为我们生活中不可或缺的一部分.同时,多媒体内容也出现了各种相关概念,如表示是否有趣吸引人的趣味性[1–4]、表示内容的是否具有美学价值

* 基金项目: 国家自然科学基金(61772535); 北京市自然科学基金(4192028); 国家重点研发计划(2016YFB1001202)

Foundation item: National Natural Science Foundation of China (61772535); Beijing Natural Science Foundation (4192028); National Key Research and Development Plan ,China (2016YFB1001202)

本文由“多媒体内容的多维度相似性计算与搜索”专题特约编辑蒋树强研究员、刘青山教授、孙立峰教授、李波教授推荐.

收稿时间: 2019-06-07; 修改时间: 2019-07-11; 采用时间: 2019-09-17; jos 在线出版时间: 2020-01-13

CNKI 网络优先出版: 2020-01-14 11:26:23, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.1125.018.html

的美学性[5,6]以及表示内容在社交网络上流行程度的流行性[7,8]等.

记忆度是一个衡量媒体数据能够被人记住的程度性指标.记忆度在一定程度上受主观因素的影响,但有研究表明,记忆度是图像的固有属性[9],所以人们在记忆上有一些共有的偏好.能够让人记住并且留下深刻记忆的媒体内容可以在我们的日常生活中为人们带来便利,为公司带来利润.令人印象深刻的广告可以帮助公司推销他们的产品,而风趣而又好记易学的视频课件可以提高课堂教学效率.媒体的记忆度预测任务旨在预测哪些类型的媒体内容对人们来说是容易记住的,它具有广泛的应用,如视频检索、视频推荐、广告设计或是教育系统.随着实时流媒体和用户生成视频井喷式的快速发展,媒体记忆度预测任务变得越来越重要,并且越来越多的研究者开始关注这个话题及其相关工作.因此,记忆度是理解媒体内容的一种方式,也是未来媒体产业发展的一个关键点.

通常来说,记忆度的预测有 3 种预测模式:粗粒度分类、回归和排序.粗粒度分类是直接预测数据样本的类别,我们预先定义一系列的记忆度分级(例如,高中低三级的记忆度),并以此为数据标签进行训练和预测.由于分类的粒度较粗,在目前的研究中,大多数工作都集中在后两种格模式.回归的模式与其他回归任务相同,它意味着用回归模型直接计算记忆度分数.排序的模式与回归有所不同.在排序的模式中,我们利用排序模型直接对样本列表进行排序,同时也可以得到记忆度的分值.

在现有的记忆度预测工作中,无论采取哪一种预测模式,人们都把注意力放在一个整体的内容表示上.例如,我们拿到一个视频的描述性标题,就可以把它向量化为特征作为视频的一个表示;对于视频帧,我们可以利用 SIFT(scale invariant feature transform,尺度不变特征变换)、LBP(local binary pattern,局部二值模式)等人工设计的特征对其进行一个整体的表示,或是用 CNN(convolutional neural network,卷积神经网络)直接学习到一个特征.这样做能够快速达到预测记忆度的效果,但却没有考虑到一些局部的信息,例如,物体的位置分布、视频中不同物体的特征表示等等.所以在本文中,我们会从全局和局部的不同角度来对视频的记忆度进行预测.

本文的主要贡献包括:

(1) 从全局和局部的不同视角对视频记忆度的预测进行探索,从全局角度分别研究了视频的描述性文本和视频中物体的空间分布状态的表示能力,从局部角度研究了视频中物体的表示.

(2) 本文提出了一种基于全局上下文表示模块、空间布局表示模块和局部物体注意力模块的视频记忆度预测模型,该模型从 3 个不同角度对视频进行特征表示,最后融合,可以把模型的预测能力进一步提升.

(3) 在 MediaEval 2018 Media Memorability Prediction Task 的数据集上验证了模型的有效性.数据集包含 8 000 个视频,并且视频的内容和场景十分丰富,全部为自然场景而非实验室人为采集,所以贴近生活和自然的数据更能说明模型的可靠性.

本文第 1 节总结和介绍记忆度预测的相关工作.第 2 节对提出的记忆度预测模型的总体框架以及细节进行阐述.第 3 节介绍记忆度预测的实验结果并加以分析.最后第 4 节总结全文并对未来的研究方向进行初步设想与探讨.

1 相关工作

1.1 记忆度预测任务模式

有的工作将图像的记忆度预测任务定义为一个分类任务[10],这些研究者使用了支持向量回归器(support vector regression)、*k*-最近邻(*k*-nearest neighbor)算法和决策树(decision tree)这 3 个回归模型作为基线,并用以 SoftMax 为最终分类层的多层感知器(multilayer perceptron)进行预测.通过设置一些阈值,将数据集提供的记忆度分数转换为 k 类标签.这样做的好处是可以把记忆度预测任务以一个统一分级标准重新定义(例如分为高中低,或分为 10 个不同等级),能够很快速地应用到不同的分类模型上去,并且可以根据需要自定义不同的记忆度粒度.但是,如果想要更加细粒度和精确的结果,还是需要用连续值而非离散值去表示记忆度.通过直接对记忆度标签进行回归预测,这是最直接也最简单而有效的一种方式[11].利用排序的思想进行记忆度的计算,优势在于

可以利用到不同样本之间的关系来进行学习.例如,在成对排序中利用正负样例的记忆度差别关系进行学习.有的工作使用了两种排序模型,即成对排序和马尔可夫决策过程排序[12].通过对排序模型的训练,最终模型可以根据输入视频得到一些内部评价得分以表示视频的兴趣度,同样的计算方式也可以用在记忆度的预测中.

1.2 图片记忆度预测

在多媒体的记忆度预测任务中,目前研究者主要集中在图片和视频领域上,而图片是目前记忆度预测的最大"热土"[13−15].许多研究者利用卷积神经网络自动学习到视觉上的特征表示,同时,如果图片有对应的描述性标题,也会被拿来用此次向量进行嵌入表示,最后融合成为多模态的整体特征[13].有研究者用于训练好的神经网络模型并进行微调来预测图片的记忆度,并同时引入情感计算的结果,分析了记忆度与情感空间之间的关系[14].有些工作探索了一些基于深度学习的图像特征,并分析了高、低记忆度图像中的模式[15].在已有研究中,文献[16]用长短时记忆网络(long short-term memory)并结合注意力机制来进行图片记忆度的预测,把所有时间不得输出结果综合计算出最终的记忆度[15].

1.3 视频记忆度预测

视频的记忆度与图像有所不同.直观地来看,静态对象和场景所显示的信息是有限的,所有扩展元素都应该由人自我探索和想象.例如,一张插着蜡烛蛋糕的图片,旁边还有一个小朋友,我们可能会想象这是小朋友在过生日,也许之后她还会吹灭蜡烛,许愿,唱生日歌等等.但视频中的动态因素可以直观地讲述故事,丰富的内容包含更多的元素,可以直接提取和利用.视频包含丰富的时序信息,但目前对于视频的记忆度研究还很少,对于时序信息的利用可借鉴其他视频任务,如视频分类等.从特征上来讲,我们可以用于对训练好的模型进行微调和特征抽取,如 C3D(convolutional 3D,3D 卷积)、HMP(histograms of motion patterns,运动模式直方图)等;从模型上来讲,循环神经网络(recurrent neural network)能够很好地捕捉视频内的时序信息.Shekhar 等人设计了 Show and Recall 的标注方法,在 TRECVID 2012 上进行了标注,并用随机森林(random forest)模型在一些特征上进行了测试,例如 C3D、图像显著性、色彩特征等等[17].

1.4 待解决问题

目前,人们对记忆度预测的任务越来越关注,一些研究开始尝试构建视频记忆度数据集,并设计一些简单的基线方法.然而,大家对于记忆度仍然没有一个共同的定义.标注的准则、评估指标等等都仍是一个开放的问题.这些问题同样也是机遇,例如一些挑战和比赛的数据集为促进这项任务研究的发展提供了更多的机会[1,9,18].

2 视频记忆度预测模型

之前一些研究表明,视觉输入至少以 3 种不同的方式在人类记忆中表示:详细的感官表示、语言文字性描述的表示和视觉结构的示意性表示[11].假设当我们观看一段视频时,将在记忆中构建 3 种信息:全局上下文表述性信息、空间布局信息和局部对象的信息.描述性标题是一种很好的全局上下文信息.当一个人观看一段视频时,在记忆中会产生一个语言描述,它是整个过程的概括,决定了语义内容的骨架.空间布局是对视频中检测到的对象的模式和结构的抽象表达.当人们开始观看视频时,空间布局是视频的一种非常直观的表达.平面视图、开放空间、对称性,这些不同的属性可能会对人们的记忆产生潜在的影响.接下来,在观看视频的过程中,人们可能会将注意力转移到不同的物体上.很自然地,我们可以设计一种模拟注意力转移的方法以捕捉人们是如何在全局语义信息的指导下使用他们的注意力的.

研究发现,视频的描述性标题在用词向量进行嵌入后有很好的特征表示能力,用于记忆度预测有很好的效果[15],所以我们可以用数据的标题进行词向量潜入表示,作为视频的全局语义信息.无论是在图像还是视频记忆度预测任务中,之前的研究工作很少关注到图像或视频中的各种物体.在之前的一项工作中,研究者设计了一个游戏,并通过将收集来的数据进行分析后发现,物体的数量和类别对于图片的记忆度是有影响的[18].利用 LSTM(long short-term memory,长短时记忆网络)和注意力机制可以捕捉人们关注的区域[2],但它没有关注到对象级别.有很多视觉因素能够影响对象记忆度,如颜色和对象类别.还有一些工作研究了物体与图像记忆

度之间的关系,发现图像中记忆度高的物体对图像整体的记忆度有很大的影响[19].这项工作很好地解释了物体和图像记忆能力之间的关系,但并没有说明人们对每个物体的关注程度.因此,我们提出了利用注意力机制来学习包含不同对象的单个帧的表示,这样能够考虑到帧中的所有对象.而且,对象和场景的语义信息(如场景类别)是记忆度的主要基础[20].因此我们进一步提出对象的空间布局,并通过一个简单的 CNN 来学习其特征表示.至此,我们得到了 3 个不同的模块组成最终的记忆度预测模型.第 1 个是全局性的上下文表示模块,它用描述性视频标题的词向量嵌入来表示全局的上下文语义.第 2 个是空间布局表示模块,它检测出帧中的不同物体的边界框并生成一个空间布局的模板,之后,CNN 学习到一个空间布局的向量表示.最后是局部的物体注意力模块,它通过注意力机制将帧中的不同物体赋予不同的权重,利用所有局部的物体信息得到一个视频帧的整体特征表示.

综上所述,本文提出了一种基于全局上下文表示模块、空间布局表示模块和局部对象注意力模块的视频记忆度预测模型.图 1 所示为所提出模型的大致框架.给定一个视频和相应的标题,我们首先可以得到标题的嵌入表示,即全局上下文表示,这是对视频的全局和一般性描述.然后通过一个预先训练的 Faster-RCNN(faster region-based convolutional network)获得每个目标的目标特征和边界框.我们用阴影填充边界框中的区域,即用黑色方块代替物体,白色填充非物体的背景,得到一个对象的空间布局,然后将其输入一个简单的 CNN 网络,转换为一个特征向量表示.对于局部对象,我们利用全局上下文,即标题的嵌入特征作为注意力机制中的 query 来计算在一个帧中检测到的每个对象的权重,对象特征的加权平均和就是该帧的最终特征表示.视频的所有帧表示都将是 GRU(gated recurrent unit)网络的输入.最后,我们利用后期融合,即将这 3 个部分的平均分数作为最终的视频记忆度分数.

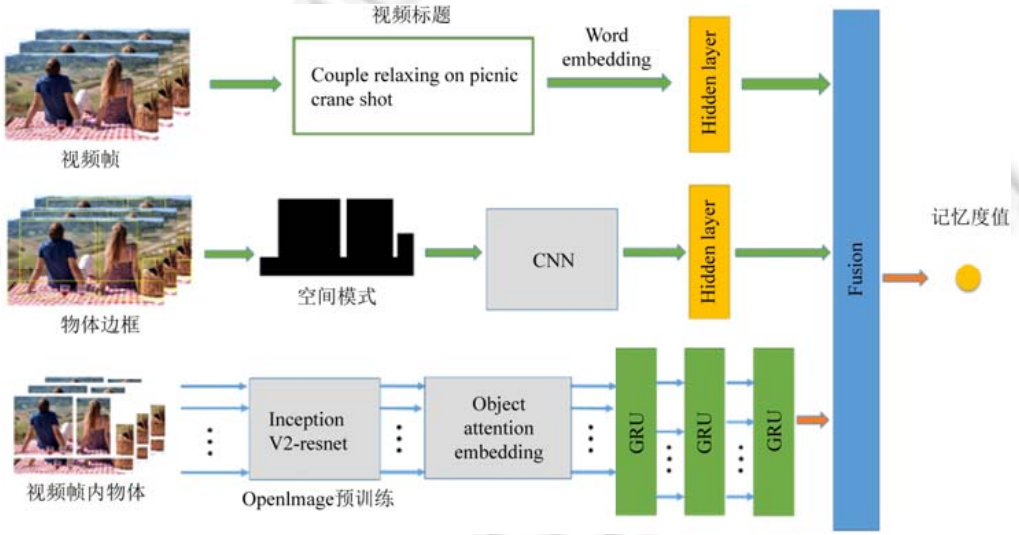

Fig.1 Video memorability prediction model based on global and local information

图 1 基于全局和局部信息的视频记忆度预测模型框架

2.1 全局性上下文表示模块

视频对应的描述性标题具有天然的概括性,能够从全局的角度总结一段视频的内容.当然,内容的丰富性或多或少会因为标题的具体描述信息而有所改变.如图 2 所示的第 1 个例子,它可以被简单地描述为“人们在野餐”,也可以被详细地描述为“一对情侣坐在野餐垫上眺望远方”.但无论是哪种描述,它都能作为一个全局性的语义总结,不会偏离视频内容.

对于单个视频来说,我们把对应标题进行分词,去除停用词,自动纠错,最后计算所有词语的 GloVe(global vectors for word representation,全局词向量表示)词向量嵌入,然后将它们的平均值作为句子级特征来表示上下

文信息.

Fig.2 Some samples which include a frame extracted from video, spatial layout, corresponding caption and long-term and short-term memorability ground-truth

图 2 数据样例:从视频中截取的一帧,对应标题、空间布局、视频长时记忆度和短时记忆度

2.2 空间布局表示模块

直观来看,图像或视频中的空间布局有两条线索,即物体之间的排列规则和空间分布.对于物体之间的排列规则,我们可以想象一下:当人们拍照时,他们通常站得井然有序,照片看起来干净整齐,或是阅兵式的仪仗队整齐划一.在空间分布上,摄影师通常遵循一些摄影技巧和规则,如黄金分割法.因此,在记忆度预测中,我们认为并假设物体的空间布局是非常重要的.为了探索视频的空间布局,我们设计了一个简单描述视频空间布局的模板.首先用 Faster-RCNN 检测视频帧中的对象,并得到相应的特征和对应边框.然后用值 1 填充对象的像素,用值 0 填充其余的像素.最后,我们可以得到一个由 0 和 1 组成的掩模,其中物体整体被黑色像素代替,而非物体背景被白色像素代替.图 3 给出了计算给定帧的掩码的过程.

Algorithm 1. 空间布局掩模.

```
Input: X(视频帧像素矩阵), H(视频帧高度),
       C(视频帧宽度)
Output: M(掩模矩阵)
所有物体边框bboxes=ObjectDetection(X)
for i from 1 to H do
    for j from 1 to W do
        M[i][j]=0;
        for bbox in bboxes do
            if X[i][j] in bboxes do
                M[i][j]=1;
                break;
            end
        end
    end
end
```

Fig.3 Spatial layout mask algorithm

图 3 空间布局掩模算法

最后,我们把得到的视频内所有掩模进行平均,并输入到一个简单的 CNN 中,其结构为:卷积层 Conv1 包含 30 个 5×5 的卷积核,步长为 1;Maxpooling 层 maxpool 1,pool size 为 2×2,步长为 2;卷积层 Conv2 包含 15 个 3×3 的卷积核,步长为 1;Maxpooling 层 maxpool 2,pool size 为 2×2,步长为 2;全连接层 hidden_layer 1 维度为 512;dropout 层,dropout rate 为 0.5;全连接层输出维度为 1,表示最终记忆度值.

2.3 局部物体注意力模块

在观看一段视频时,人们的注意力是在不断转换的,也许我们把目光一直锁定在某个物体上,抑或是在不同的物体间来回切换.很多情况下,我们第一眼会看到最大的物体,所以一种简单的方法是使用最大对象来表示帧.但是,它缺少来自其他对象的补充信息.例如,在两帧中检测到的最大对象都是人,一帧位于海滩上,周围是度假者,另一帧则位于办公室,周围是桌子和打印机.环境和周围的物体对语义有很大的影响.

Soft Attention 机制首次应用于机器翻译中.Soft Attention 为每个元素生成概率权重,我们可以利用它为视频帧中的每个对象赋予不同的注意力权值.为了充分利用隐藏在所有对象中的信息,我们提出利用 Soft Attention 将框架中的所有对象嵌入到单个表示中.将第 2.1 节中提到的标题嵌入特征作为一个 query,并得到每个帧中所有对象功能的加权平均值.最后利用 3 层 GRU 网络对时间信息进行捕获,得到一个记忆度得分.

如果标题由 n 个单词组成,可将标题中的第 i 个单词定义为 W_i,而 $f_{ew}(x)$表示单词 x 的词嵌入表示,则全局上下文表示如下:

$$\frac{1}{n}\sum_{i=1}^{n} f_{we}(w_i).$$

g_t 表示注意力权重 α_t 以及在第 t 帧的所有物体特征的加权和.

$$g_t = \sum_{i=1}^{M} a_{i,t} x_i \alpha \in \mathbb{R}^M x_i \in \mathbb{R}^D.$$

注意力机制的权重有网络训练确定,最终通过 softmax 函数表示为一个权重向量:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{M}\exp(e_{t,k})}.$$

未经过 softmax 前的注意力权重是由物体特征和标题的嵌入特征乘积而得来的:

$$e_{i,j} = f_{score(x_i,c)}.$$

$f_{score(x_i,c)}$ 为计算每一个物体的注意力得分的得分函数:

$$f_{score(x_i,c)} = v^{\mathrm{T}} \tanh(Wc + Ux_i + b),$$

其中, $v \in \mathbb{R}^D, W \in \mathbb{R}^{D\times C}, U \in \mathbb{R}^{D\times D}, b \in \mathbb{R}^D$ 分别为网络的权重和偏置.

对于单个视频来说,我们首先抽取视频帧,并用 Faster-RCNN 检测帧中的对象并得到对应的特征和边界框,并按照面积由大到小排列.最后,我们建立了一个 3 层 GRU 网络来捕获整个视频的时序信息以预测出一个记忆度分值.

3 实验与分析

3.1 数据与任务描述

数据集由 8 000 个短的无声视频组成,视频根据许可证共享,许可证允许在 MediaEval 2018 环境中使用和重新发布.我们首先根据视频的记忆度得分进行排名,升序降序皆可(保证记忆度值的分布),并以固定步长值 4 对视频样本进行采样,每 5 个视频采样出一个,作为测试数据.最后,将开发集分为两部分,分别是训练集 6 000 个视频和测试集 2 000 个视频.

这些视频是从专业人士在制作视频时使用的原始视频中提取出来的.每个场景的持续时间为 7s,内容丰富,包含不同的场景类型.每个视频还附带其原始标题.这些标题通常被视为文本元数据,可能有助于预测视频的记忆度.

数据集包含两种标签,即长期记忆标签和短期记忆标签,分别对应于两个子任务.

- 短期记忆度子任务:该任务包括预测给定视频剪辑的“短期”记忆度得分,这反映了观看视频几分钟后记住的可能性;
- 长期记忆度子任务:这项任务包括预测给定视频剪辑的“长期”记忆度得分,这反映了观看后 1~3 天记住的可能性.

对于这两个子任务,官方的评估指标是所有视频的真实记忆度和预测记忆度之间的 Spearman's rank correlation.

3.2 基线系统

一般来说,我们使用回归方法来预测每个视频的记忆度分数,并考虑后期融合来结合不同的特征.采用了两种融合策略,即分数平均和二层回归.

首先,我们使用不同的单一特征进行回归,得到视频的记忆度分数.为了融合多个特征,考虑了两种策略.第 1 种是对同一视频中不同类型特征的分数进行平均,得到的分数是该视频的最终记忆度分数.第 2 种是对于第 2 层回归,我们将来自同一视频的不同特征的分数作为特征连接起来,并输入第 2 层回归模型,从而预测最终的记忆度分数.

因为视频是无声的,因此我们探索了视频和文本的特征,特别是一些高级和语义的表达.视频的标题很短,只有几个字.我们认为人们可能会对某些特定的物体或它们的组合印象深刻.每个词的意义都应该嵌入句子的表示中,以便于记忆预测.

预训练嵌入包含大量语义信息,有助于对句子的语义进行编码.我们尝试用 GloVe[21]词向量作为文本特征.结合每个单词的嵌入,以不同的方式生成句子的表示.首先,简单地把它们加起来,取每个维度的平均值.之后,将平滑的 IDF(inverse document frequency,逆文本频率指数)作为每个单词的权重[22].然后,尝试预先训练的 skip-thought[23]模型.最后,我们还尝试 ConceptNet[10].通过这 4 种方法,可以获得不同类型的视频级表示.

对于视觉特征,我们考虑一些神经网络学的特征和美学特征,包括 C3D[24]、HMP[25]比较、I3D[26]、美学[27].在基线系统,采用 C3D、HMP 和美学特征.此外,我们还提取了 I3D 中 RGB 分支倒数第 2 层的特征.

我们用两种回归器作为基线模型,即支持向量回归(SVR)和随机森林回归(RFR).参数由网格搜索确定.SVR 中的惩罚参数 C 从 0.125~32.内部评估器数量的搜索范围是[100,1000],步长为 100;最大深度范围是[2,10],其中,步长为 2.I3D 模型在 ImageNet 和 Kinetics 进行了预训练.

3.3 模型表现

长时和短时记忆度预测的不同特征结果分别如图 4 和图 5 所示.从图 4 图 5 中可知,文本特征普遍比视觉表示表现得更好.我们认为,标题中包含了关于视频元素的更清晰的描述.如果一个特定的对象是用一个词来描述的,那么嵌入这个词就可以描述这个对象和整个环境中其他对象的关系.视觉特征可能包含一些区域的细节,但不太直观,也许尚未捕捉到与记忆度相关的部分.

对于空间布局,我们考虑 3 种不同的设置,即简单遮罩、重叠遮罩和面积大小遮罩.简单遮罩即为我们用值 1 填充对象的像素,其他像素用值 0 填充.带重叠的遮罩表示重叠区域由包含此区域的对象数填充.而面积大小遮罩是用区域大小而不是值 1 填充像素.对于每种设置,我们还考虑沿时间维度融合所有帧的两种策略:平均和 LSTM.表 1 和表 2 给出不同空间布局策略的结果.可以看到,重叠在空间布局中并不重要,简单遮罩就可以很好地表示空间布局.此外,时序信息对性能没有帮助.我们认为,首先,大多数视频的场景没有剧烈的变化,捕捉到的时序信息并不能很好地和平均策略有所区分.其次,对于场景变化不大的视频来说,人们可能更关注视频的整体空间布局,而不是随时间变化的模式.如果我们将实验数据换成电影广告等视频,也许动作场面的设计、场景的变幻更能抓住人的眼球.

表 3 为我们所提方法中不同策略的结果.可以发现,将所有帧的特征取平均结果即可比基线系统中的最佳结果表现得更好.用 GRU 进行时序信息的捕捉,可以进一步提高效果.在此基础上,我们又加入了物体的注意力

机制,捕捉物体之间的权重分配信息,从而进一步提高模型整体的表现.由此可知,记忆度预测可以从全局语义信息、空间分布、时序信息以及局部的物体信息中受益.

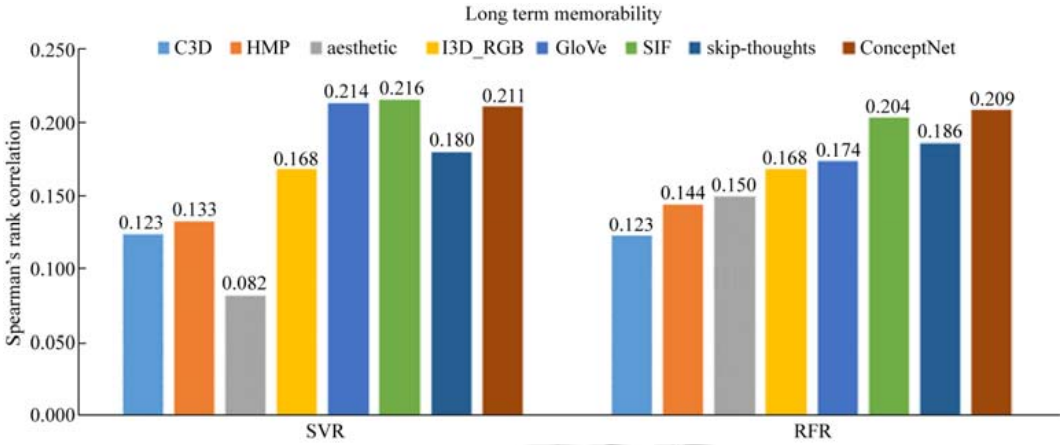

Fig.4 Results of different features for long-term memorability on the local test set

图 4 不同特征在本地测试集上长时记忆度的结果

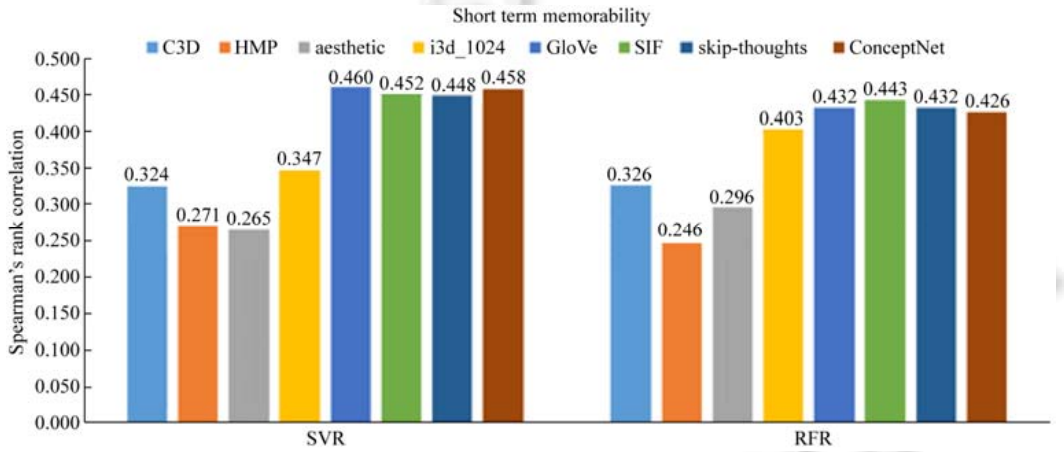

Fig.5 Results of different features for short-term memorability on the local test set

图 5 不同特征在本地测试集上短时记忆度的结果

Table 1 Results of spatial layout for long-term memorability

表 1 空间布局模块的长时记忆度结果

策略	简单遮罩	重叠遮罩	面积大小遮罩
视频帧平均	0.144 0	0.129 0	0.141 1
LSTM	0.138 8	0.121 5	0.138 4

Table 2 Results of spatial layout for short-term memorability

表 2 空间布局模块的短时记忆度结果

策略	简单遮罩	重叠遮罩	面积大小遮罩
视频帧平均	0.280 3	0.270 5	0.277 3
LSTM	0.265 4	0.262 3	0.260 5

Table 3 Results of two strategies in object branch on the test set
表 3 两种处理物体特征策略在测试集上的结果

策略	物体特征平均	物体特征时序	物体注意力
Long-term(无空间布局模块)	0.224 8	0.228 1	0.234 0
Long-term(空间布局)	0.231 2	0.238 5	0.241 6
Short-term(无空间布局模块)	0.447 1	0.458 0	0.465 2
Short-term(空间布局)	0.469 5	0.470 2	0.471 1

从表 1 和表 2 对比以及表 3 的内部对比来看,短时记忆度比长时记忆度的可预测性要高很多.原因在于短时记忆度所表示的是测试者在几分钟内能记住该视频的程度,而长时记忆度的测试时间是 1~3 天后.因此这会造成两个现象,一是从数据的标注来看,短时记忆度普遍高于长期记忆度,因为人短期内能够较清晰地记住几分钟前所看的视频;二是从实验结果来看,短期记忆度更好预测.从图 4 和图 5 来看,记忆度长短时对不同特征的记忆度预测能力影响不大,也就是说,在长时预测较好的特征普遍在短时记忆度上也有良好表现.当然会有一些例外,如 C3D 特征在短时记忆度上的相对预测能力要高于长时记忆度,我们未来工作中也会分析这些原因,比如是否是因为时序上的信息更能影响人的短期记忆而长期记忆更受全局的视频表征影响.

我们挑选了一些视频样例进行分析,发现其中一些视频描绘了物体的具体部位或是特写镜头,而其中一些视频显示了整体的场景,如自然景观、一些人物的故事.

图 6 展示了一些数据样例.图 6(a)和图 6(b)所示长短时记忆度都很高;图 6(c)和图 6(d)所示长时记忆度很高而短时记忆度很低;图 6(e)~图 6(h)所示都有着较低的长时记忆度和较高的短时记忆度.

通过观察很多视频样例及其标签后,我们得出如下一些猜想.

- 短期标签较低的视频通常具有长期标签较低的特征;
- 具有较高短期标签和较低长期标签的视频通常描述一些特写镜头或一些静态的常见对象和场景;
- 有少量的视频具有较低短期标签和较高长期标签.这些视频通常有开放和广阔的场景;
- 一些有趣的物体或场景可以使长期和短期记忆度得分很高,例如一个戴潜水镜的男人坐在海滩上用笔记本电脑工作,样品如图 6(a)所示.因此,视觉和文本内容背后的语义信息是一个值得探讨的重要问题.

从这些样例可以看出,如果一个视频在长期内是值得纪念的,那么在短期内通常也是值得纪念的.相反,具有较高的短期记忆度视频无法决定长期记忆度.所以,可以将对这些样本的分析结论作为一个猜想,推动后面的研究和分析.

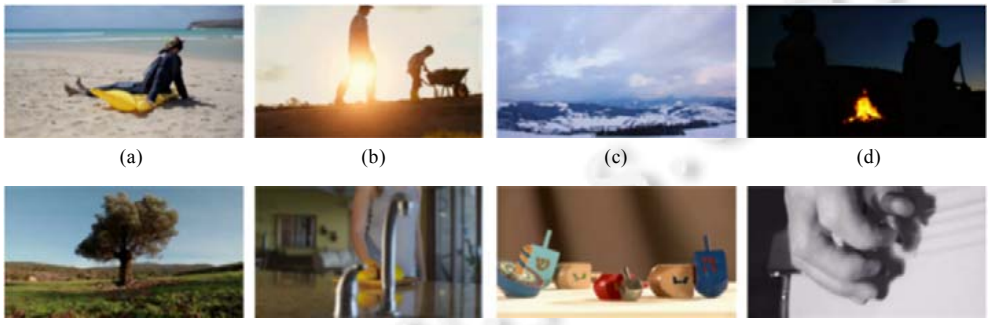

Fig.6 Some samples of the dataset
图 6 数据集中的一些样例

4 总结与展望

互联网、移动设备以及软件服务等不同因素的共同发展,使得互联网上的视频也发生爆炸式的增长.精确预测视频的记忆度可能会给人们的生活带来更大的便利,也能够给企业带来大量商机与发展,例如多媒体检索

与推荐、教育系统、广告设计等等.在本文中,我们从全局和局部两个角度探索了视频的视觉和文本的特征表示,并提出一个视频记忆度预测模型.实验结果表明,在全局的表示上,文本表示的性能优于视觉特征.空间布局特征的表现也十分良好,甚至比一些深度神经网络的特征表示更加有力.同时,局部的物体注意力机制的学习也能很好地捕捉到一些记忆度的信息.对不同对象使用注意力机制,可以有效地将所有对象嵌入到一个单一的表示中,并显示出显著的性能提升.在未来的工作中,我们将重点关注视觉语义的表达和视频是对象与记忆度关系的相关工作,如探索不同对象之间的关系,设计更稳定的模型来预测视频的记忆度.

References:

[1] Romain C, Claire-Hélène D, Duong NQK, Sjöberg M, Ionescu B, Do TT, Rennes F. In: Proc. of the MediaEval 2018: Predicting Media Memorability Task. Sophia Antipolis, 2018. 29–31.

[2] Fajtl J, Argyriou V, Monekosso D, Remagnino P. AMNet: Memorability estimation with attention. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6363–6372. [doi: 10.1109/CVPR.2018.00666]

[3] Gygli M, Grabner H, Riemenschneider H, Nater F, Van Gool L. The interestingness of images. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 1633–1640. [doi: 10.1109/ICCV.2013.205]

[4] Zhong ZM, Guan Y, Hu Y, Li CH. Mining user interests on microblog based on profile and content. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):278–291 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5030.htm [doi: 10.13328/j.cnki.jos.005030]

[5] Bhattacharya S, Sukthankar R, Shah M. A frame-work for photo-quality assessment and enhancement based on visual aesthetics. In: Proc. of the ACM Int'l Conf. on Multimedia. 2010. 271–280. [doi: 10.1145/1873951.1873990]

[6] Wang CH, Pu YY, Xu D, Zhu J, Tao ZE. Evaluating aesthetics quality in portrait photos. Ruan Jian Xue Bao/Journal of Software, 2015,26(Suppl.(2)):20−28 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15012.htm

[7] Khosla A, Das Sarma A, Hamid R. What makes an image popular. In: Proc. of the Int'l Conf. on World Wide Web. 2014. 867–876. [doi: 10.1145/2566486.2567996]

[8] Kong QC, Mao WJ. Predicting popularity of forum threads based on dynamic evolution. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2767–2776 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4730.html. [doi: 10.13328/j.cnki.jos.004730]

[9] Isola P, Xiao JX, Parikh D, Torralba A, Oliva A. What makes a photograph memorable. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2014,36(7):1469–1482. [doi: 10.1109/TPAMI.2013.200]

[10] Speer R, Chin J, Havasi C. ConceptNet 5.5: An openmultilingual graph of general knowledge. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco, 2017. 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[11] Phillips WA. On the distinction between sensory storage and short-termvisual memory. Perception & Psychophysics, 1974,16(2):283–290. [doi: 10.3758/BF03203943]

[12] Wang S, Chen SZ, Zhao JM, Jin Q. Video interestingness prediction based on ranking model. In: Proc. of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-ModalAffective Computing of Large-Scale Multimedia Data. ACM, 2018. 55–61. [doi: 10.1145/3267935.3267952]

[13] Squalli-Houssaini H, Duong NQK, Gwenaëlle M, Demarty CH. Deep learning for predicting image memorability. In: Proc. of the 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. 2371–2375. [doi: 10.1109/ICASSP.2018.8462292]

[14] Baveye Y, Cohendet R, Da Silva MP, LeCallet P. Deep learning for image memorability prediction: The EmotionalBias. In: Proc. of the ACM on Multimedia Conf. 2016. 491–495. [doi: 10.1145/2964284.2967269]

[15] Zarezadeh S, Rezaeian M, Sadeghi MT. Image memorability prediction using deep features. In: Proc. of the Electrical Engineering. 2017. 2176–2181. [doi: 10.1109/IranianCEE.2017.7985423]

[16] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Computer Science, 2014.

[17] Shekhar S, Singal D, Singh H, Kedia M, Shetty A. Show and recall: Learning what makes videos memorable. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision Workshops (ICCVW). Venice: IEEE, 2017. 2730–2739. [doi: 10.1109/ICCVW.2017.321]

[18] Demarty CH, Sjöberg M, Ionescu B, Do TT, Gygli M, Duong NQK. In: Proc. of the Mediaeval 2017 Predicting Media Interesting Nesstask. 2017.

[19] Dubey R, Peterson J, Khosla A, Yang M, Ghanem B. What makes an object memorable. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). 2015. 1089–1097. https://doi.org/10.1109/ICCV.2015.130

[20] Isola P, Xiao J, Torralba A, Oliva A. What makes an image memorable. In: Proc. of the CVPR. 2011. 145–152. https://doi.org/10.1109/CVPR.2011.5995721

[21] Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: Proc. of the Empirical Methods in Natural Language Processing (EMNLP). 2014. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[22] Arora S, Liang YY, Ma TY. A simple but tough-to-beat baseline for sentence embeddings. In: Proc. of the ICLR. 2017.

[23] Kiros R, Zhu YK, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S. Skip-thought vectors. In: Advances in Neural Information Processing Systems 28. Curran Associates, Inc., 2015. 3294–3302.

[24] Du T, Bourdev L, Fergus R, Torresani L. Learning spatio temporal features with 3D convolutional networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4489–4497. [doi: 10.1109/ICCV.2015.510]

[25] Almeida J, Leite NJ, Torres RDS. Comparison of video sequences with histograms of motion patterns. In: Proc. of the IEEE Int'l Conf. on Image Processing. 2011. 3673–3676. [doi: 10.1109/ICIP.2011.6116516]

[26] Carreira J, Zisserman A. Quo Vadis, action recognition? A new modeland the kinetics dataset. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017. 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[27] Haas AF, Guibert M, Foerschner A, Co T, Calhoun S, George E, Hatay M, Dinsdale E, Sandin SA, Smith JE, Vermeij MJA, Felts B, Dustan P, Salamon P, Rohwer F. Can we measure beauty? Computational evaluation of coral reef aesthetics. 2015. https://doi.org/10.7717/peerj.1390

附中文参考文献:

[4] 仲兆满,管燕,胡云,李存华.基于背景和内容的微博用户兴趣挖掘.软件学报,2017,28(2):278–291. http://www.jos.org.cn/1000-9825/5030.htm [doi: 10.13328/j.cnki.jos.005030]

[6] 王朝晖,普园媛,徐丹,祝娟,陶则恩.人像照片的美感质量评价.软件学报,2015,26(Suppl.(2)):20−28. http://www.jos.org.cn/1000-9825/15012.htm

[8] 孔庆超,毛文吉.基于动态演化的讨论帖流行度预测.软件学报,2014,25(12):2767–2776. http://www.jos.org.cn/1000-9825/4730.html [doi: 10.13328/j.cnki.jos.004730]

王帅(1993－),男,学士,CCF 学生会员,主要研究领域为情感计算.

陈师哲(1994－),女,学士,CCF 学生会员,主要研究领域为多模态内容理解.

王维莹(1996－),女,学士,主要研究领域为多媒体计算.

金琴(1972－),女,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为多媒体语义理解,情感计算.