文章编号：2095-2163(2020)10-0001-05 中图分类号：TP18 文献标志码：A

基于差异特征的强化学习视频摘要

李巧凤，赵 烨
（合肥工业大学 计算机与信息学院，合肥 230009）

摘 要：视频摘要是对视频内容的高度概括，选择出具有多样性和重要性的视频帧子集。文章从关键帧代表性不够全面的角度出发，提出一种利用多路特征来提取视频关键帧的方法，通过卷积神经网络（convolutional neural network，CNN）和长短时记忆网络（long short-term memory，LSTM）来预测视频帧被选中的概率。将提取出的视频帧的原始特征送入 LSTM，将处理过的两两视频帧特征的差特征也做同样的处理，差特征包含了相邻视频帧之间更多不同的信息。由于 LSTM 长期依赖的特性，使得整个网络可以学习到视频上下文之间更多的信息，通过对处理过的两路特征做得分融合，作为判断视频帧被选择与否的最终得分。文中的强化学习机制对视频摘要有优化的作用，实验在两个基准数据集 SuMme 和 TVSum 上进行。结果表明，该方法能够显著提高调和平均数（F-score）指标。

关键词：视频摘要；多路特征；卷积神经网络（CNN）；长短时记忆网络（LSTM）；得分融合

Video Summary of reinforcement learning video based on difference features

LI Qiaofeng, ZHAO Ye
（School of Computer and Information, Hefei University of Technology, Hefei 230009, China）

【Abstract】 The video summary is a highly concise video content that extracts a subset of video frames of diversity and importance. From the perspective of improving the quality of video summary, this paper proposes a method to extract video key frames by using multi-channel features, which is through convolutional neural network (CNN) and long-term memory network (LSTM) to predict the probability that a video frame will be selected. The method is to send the original symbols of the extracted video frames to the LSTM, and on the other hand, the difference features of the processed two-two video frame features are also treated the same, and the difference features include more different between adjacent video frames. Information, and because of the long-term dependence of LSTM, allows the entire network to learn more information between video contexts, by making a score fusion of the processed two-way features as the final score of whether the video frame is selected or not. The intensive learning mechanism in this paper has an optimized effect on video. We conducted experiments on two benchmark datasets, SuMme and TVSum. The results show that this method can significantly improve the F-score index.

【Key words】 video summary; multi-features; convolutional neural network (CNN); long short-term memory (LSTM); score fusion

0 引 言

作为主要的多媒体信息载体，海量的视频丰富了我们的生活，也为科技的快速发展带来巨大机遇，但每天拍摄上传到网络上的海量视频，使得视频分析耗费大量财力。因此，迫切需要更加有效的信息组织，总结和分析的技术，了解在这个视频中哪个帧和镜头是不可或缺的。本文的目的就是通过强化学习机制处理包含差异特征的多路特征，来获取得分最高的视频帧子集，即所需的关键帧。考虑到基于差异特征获取关键帧的做法相对较少，本文的思路为视频摘要的获取方法做了补充，具有一定的学术价值，通过视频关键帧提取和视频结构化生成一个有着现实意义并且能够体现视频主要内容的结构大纲。有文献提出通过对大量视频搜索与检索可以满足对所需内容的有效需求[1]，但是并没有提供实际视频内容的具体意义，很难快速找到所需的内容，基于内容频率或非冗余虽然简单有效，但是却与视频的实时性缺少直接的联系；有文献提出的预测模型用于预定义的一组类别，还有将搜寻对象限制在有限的对象域将关键帧提取的效果进行了改进[2]；有研究提出使用基于标题的图像搜索结果来寻找视频中的关键帧和重要镜头[3]。因为视频标题是通过挑选以最大程度的描述其主题，使用标题搜索的图像可以包含噪声（与视频内容无关的图像）和方差（不同主题的图像），其开发了共同原型分析技术，通过两个数据集 TVSum 和 SumMe 的联合因子

基金项目：国家自然科学基金（61876056，61502138）。

作者简介：李巧凤（1989-），女，硕士研究生，主要研究方向：智能信息处理；赵 烨（1975-），女，博士，副教授，硕士生导师，主要研究方向：机器视觉、多媒体信息处理。

收稿日期：2020-05-21

学习图像和视频之间共享的规范视觉概念,将视频摘要的质量极大地提高;有研究提出利用视觉线索、语义线索和上下文线索标签重要性预测模型[4]。采用结构支持向量机(structured support vector machine ,SSVM)公式,保证预测模型的有效训练。然后,利用正则相关分析(canonical correlation analysis,CCA)学习图像视觉特征与标签重要性之间的关系,获得鲁棒检索性能。深度视频摘要模型(deep summarization network , DSN),利用奖励机制提取视频摘要。逐渐出现的深层语意嵌入的视频摘要。提出了一种新的基于深度语义嵌入(DSSE)的视频摘要生成模型,该模型充分利用了视频摘要的边信息(标题,查询,描述),通过交互最小化两个单模态自编码器的语义相关损失和特征重构损失,可以更完整地学习视频帧与边信息之间的公共信息。有文献基于深度神经网络的软计算技术集成在一个两层的框架中来实现多视点视频摘要(MVS)[5]。主要的思路是首先在线层执行基于目标外观的镜头分割,并将其存储在一个查找表中,该查找表将被传输到云以进行进一步处理。第二层从查找表中序列的每一帧中提取深度特征,并将其传递给深度双向长短时记忆(DB-LSTM),获取信息量的概率,生成摘要;有文献引入视频帧的空间流 RGB 图像和空间流多帧运动矢量以及输入视频的时间信息来进行视频摘要[6];有文献提出一种基于聚类的多尺度以自我为中心的视频摘要与动作排序算法[7],可以一次运行中生成多个摘要,然后再以以自我为中心的视频中出现的行为动作进行优先级的排序来获得视频摘要。如何使提取的视频摘要质量高,多样性强。从这一问题入手,文章提出采用多路特征的卷积神经网络模型来优化选取视频帧子集的质量。

1 多路特征的方法

利用差异特征进行视频中关键帧的选取,本文提出了一种多路特征进行视频关键帧提取的架构。

1.1 动机

本文提出的包含差异特征的多路特征检测关键帧的方法包含了视频帧更多的特征信息。在提出的检测方法中,既着眼于提取出的视频的原特征,又侧重于处理两两视频帧之间显著性差异对象之间的差特征,通过使用多路特征网络模型来对视频进行关键帧的检测和选取,本文的多路视频摘要网络模型如图 1 所示。

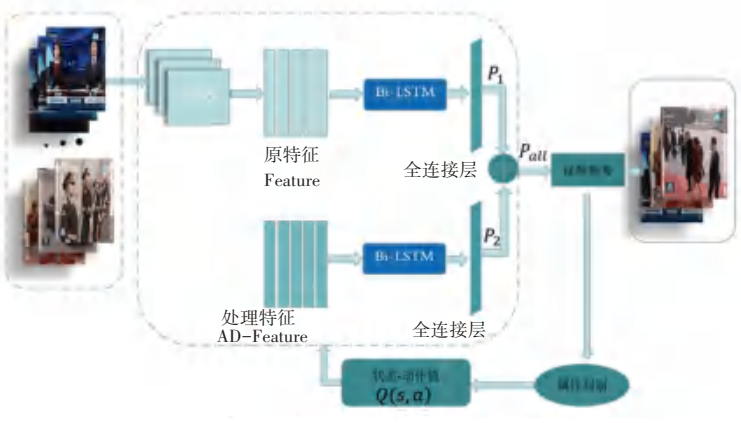

图 1 多路特征视频摘要网络

Fig. 1 Multi-channel feature video summary network

1.2 强化学习

强化学习是一种自学习系统,主要通过反复试验来学习,最终找到规律,达到学习的目的[8]。其关键要素为:智能体(agent)、环境(environment)、奖励(reward)、动作(action) 和状态(state),通过这些要素建立一个强化学习的模型,基本原理是:agent 的某个行为策略导致环境正的奖赏增大,那么 agent 以后产生这个行为策略的趋势便会增强,agent 的目标是在每个离散状态发现最优策略以使期望的折扣奖赏和最大。强化学习把学习看作试探评价过程,agent 选择一个动作作用于 environment,environment 接受该动作后状态发生变化,同时产生一个强化信号(奖或罚)反馈给 agent,agent 根据强化信号和环境当前 state 再选择下一个 action,选择的原则是使受到正强化 reward 的概率增大。本文将强化学习运用于视频摘要,通过判断选择关键帧的奖励的大小反过来影响采取该动作的概率。实验结果证明,在关键帧的提取上效果不错。强化学习模型如图 2

所示。

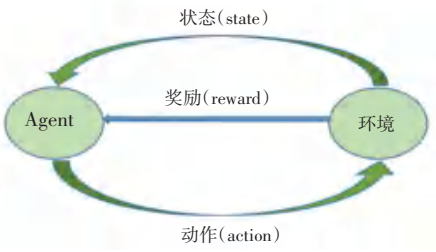

图 2　强化学习模型

Fig. 2　Reinforcement learning model

1.3　关键帧检测与提取

为了使关键帧检测模型效果好，需要一个好的关键帧解码器，本文采用端到端的编解码深度摘要网络，编码器是卷积神经网络 CNN。文中用 $x = [x_1 \quad x_2 \quad \dots \quad x_t \quad \dots \quad x_T]$ 表示视频的帧序列特征，x_t 表示表示在第 t 帧的视觉特征，解码器使用的是性能突出的双向长短时记忆网络 Bi-LSTM，把提取出的原始视觉特征 $[x_1 \quad x_2 \quad \dots \quad x_t \quad \dots \quad x_T]$ 送入 Bi-LSTM，同时用 x_{ad} 表示经过处理的帧序列特征，$x_{ad} = [x_1 - x_2 \quad x_2 - x_3 \quad \dots \quad x_t - x\,t + 1 \dots x_{T-1} - x_T]$，$x_{ad}$ 表示相邻帧序列之间差异的视觉特征，把处理后的整个视觉特征 x_{ad} 也完整地送入到 Bi-LSTM，生成 t 时刻相应的隐藏状态 h_t，Bi-LSTM 对信息的处理方式分为两个方向，前向状态和后向状态，它封装了当前帧过去的和未来的信息，经过 Bi-LSTM 处理后转换为表示向量 $v = [v_1 \quad v_2 \quad \dots \quad v_t \dots v_T]$，与 LSTM 相连接的全连接层(fully connected layer，FC)以 sigmoid 函数为每一帧预测得分，p_1 表示的是原始视频帧的重要性得分，p_2 表示的是差异视频的重要性得分。作为该视频帧是否被选择的概率，σ 代表 sigmoid 函数，通过伯努利函数 B 采取相应的动作，a_t 表示所采取的动作，$a_t = 1$ 表示第 t 帧被选取，为 0 则舍弃，式(1)~(4)，实验结果如图 3 所示。

$$v_t, h_t = BiLSTM(x_t, [h_{t-1}^{for}, h_{t-1}^{back}]), \tag{1}$$

$$P_t = \sigma(FC(h_t)). \tag{2}$$

$$P_{all} = p_1 + p_2, \tag{3}$$

$$a_t \sim B(p_{all}). \tag{4}$$

图 3　实验结果

Fig. 3　Experimental results

1.3.1　状态-动作值函数对 agent 的作用

评判视频摘要模型生成的摘要质量的高低，状态-动作值的大小就是很好的指标，由于强化学习的原理机制，状态-动作值越大，说明视频摘要生成的质量越高，这是一个不断学习的过程，以确保视频摘要的重要性和多样性。本文的模型中，重要性表示视频摘要对全文视频信息的覆盖能力，把它当做一个 k-medoids 问题，公式(5)所示：

$$E(x_t) = \min \| x_t - x_t' \|_2, \tag{5}$$

其中，t 和 t' 表示为非同一时刻，即最大重要性值可表示为式(6)：

$$Q^i = \exp[-\frac{1}{T}\sum_{t=1}^{T} E(x_t)]. \tag{6}$$

在视频摘要技术的发展过程中，已经提出了很

多衡量视频摘要多样性的模型。本文通过特征空间所选帧之间的差异大小,来评估视频摘要多样性的高低。用 $S=[f_1\ \ f_2\ \ \ldots f_t\ \ \ldots\ \ f_T]$ 表示所选的视频帧,则其两两之间的差异性,可以表示为式(7)和式(8):

$$Q^d=\frac{D(x_t,x_t^{'})}{T|T-1|},\tag{7}$$

$$D(x_t,x_t^{'})=\sum_{t\in T}\sum_{\substack{t^{'}\in T\\ t^{'}\neq t}}(1-\frac{x_t^T x_t^{'}}{\|x_t\|_2\ \|x_t^{'}\|_2}).\tag{8}$$

视频摘要的属性判别也就是关键帧的属性判别。它的属性就是重要性和多样性公式(6)的 Q^i 越高,代表重要性越强;同样道理公式(7) Q^d 越大,多样性的信息量就越丰富。

整个视频摘要用 Q^d 与 Q^i 的和状态-动作值函数 $Q(s_t,a_t)$ 表示最大奖励(reward),选择的视频帧质量越高,深度摘要网络获得的状态-动作值越大,就会促使系统选取更多这样的视频帧,二者相辅相成,式(9)。

$$Q(s,a)=Q^i+Q^d.\tag{9}$$

1.3.2 策略梯度

在不同的状态(state)采取的动作(action)也就是策略梯度 policy gradient。为了最大化状态-动作值,实验中用策略函数 π_θ 和参数 θ 来最大化期望奖励,式(10)和式(11):

$$J(\theta)=\mathrm{E}_{p_\theta(a_{1:T})}[Q(s_t,a_t)],\tag{10}$$

$$\tilde{\mathrm{N}}_\theta J(\theta)=\sum_{t=1}^{T}\mathrm{E}_{p_\theta(a_{1:T})}[\tilde{\mathrm{N}}_\theta\log\pi_\theta(a_t\mid s_t)Q(s_t,a_t)].\tag{11}$$

式中,a_t 为采取的动作;s_t 为隐藏层的状态;

$p_{\theta(a_{1:T})}$ 表示通过动作序列得到的概率分布。

为了方便计算避免个体的偏差,需要多次取样并利用均值提高其准确率,并在这里引入一个基准值 b,其为状态-动作值的平均值,则公式(11)就变形为式(12):

$$\tilde{\mathrm{N}}_\theta J(\theta)\approx\frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T}\{\tilde{\mathrm{N}}_\theta\log\pi_\theta(a_t\mid s_t)[Q_m(s_t,a_t)-b]\}.\tag{12}$$

参数 θ 的更新为公式(13):

$$\theta=\theta+\alpha\,\tilde{\mathrm{N}}_\theta[J(\theta)-\beta_1\ \|\ \frac{1}{T}\sum_{t=1}^{T}p_t-l\|^2-\beta_2\sum_{i,j}\theta_{i,j}^2].\tag{13}$$

其中,α 为学习速率;β_1 和 β_2 为平衡权重的参数,l 决定选取的视频帧的百分比。

2 实 验

2.1 数据集

本实验依然是在 2 个公共基准数据集 SumMe 和 TVSum 做评估。SumMe 数据集包含 25 个用户视频,视频的长度从 1 到 6 min 不等,记录了各种各样的事件,包括了运动、假期和烹饪等,且每个视频由 15 到 18 个人注释,每个视频有多个基本事实摘要(ground-truth)。TVSum 收录了从 YouYube 的 50 个视频,每个时长 2 到 10 min 不等,数据集覆盖 10 类别,包括了动物美容、汽车轮胎和让汽车脱困等内容,而且 TVSum 可以提供帧级重要性分数。这些都是作为基本事实标签。在实验中随机将数据集分为训练集和测试集,其中训练集占比 80%,测试集占比 20%。

2.2 评价标准

实验结果采用目前通用方式计算 F-score 来评估本文提出的方法。即量化视频摘要与 ground-truth 之间的相似性,生成的视频摘要(A)和 ground-truth (B)。精确度(P)和召回率(R)的公式定义式(14)和式(15):

$$P=\frac{A\cap B}{A},\tag{14}$$

$$R=\frac{A\cap B}{B}.\tag{15}$$

F-score 定义为式(16):

$$F=\frac{2PR}{P+R}\times 100\%.\tag{16}$$

2.3 实验细节

本文对视频进行 2 帧/秒的速度采样到帧序列中,选择使用 GooLeNet 的 pool5 层的输出,在 ImageNet 上训练,实验中 RNN 单元的隐藏状态维数为 256,epoch 最大数量为 60,达到这个数量,训练将停止。在实验中随机将数据集分为训练集和测试集,其中训练集占比 80%,测试集占比 20%。视频摘要的长度控制在原视频的 15%。学习率为 0.000 01。

2.4 实验结果分析

在 SumMe 数据集中 video_8、video_20 和 TVSum 数据集中 video_33、video_42 的实验结果如图 4 所示,红色的曲线表示真实得分(ground truth),蓝色表示的是本文的方法生成预测得分。从结果中可以得出本文方法的预测的得分与数据集中真实得分的曲线对比。通过多次的实验结果表明本文的方法预测出的分数曲线可以很好的去接近真实分数。

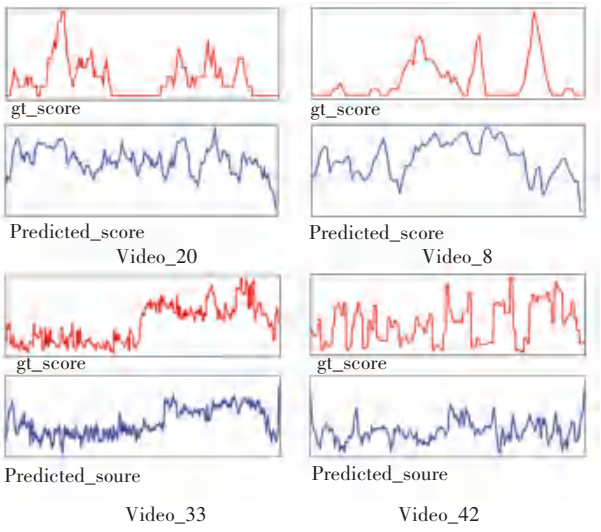

图 4　SumMe 和 TVSum 数据集实验结果

Fig. 4　SumMe and TVSum data set experiment results

本文方法在 SumMe 和 TVSum 数据集上和其他方法的做了比较。在同等的实验条件下，由表格 1 的后两项实验结果可以看出，在 SumMe 和 TVSum 两个数据集上，本文的方法比原始单一特征的 F-score分别提高了 1%和将近 2%，对于表 1 中所有的F-score值，在实际的操作中，以训练集和测试集所占的百分比为基准，表中的结果都是测试集的平均值。如在 SumMe 数据集中，计算的是 5 个视频结果的平均分为 F-score 值，在 TVSum 数据集中则是 10 个视频结果的平均分数。这充分说明了本文方法进行视频摘要的有效性。

同时，还可以看出本文的实验结果比目前绝大部分视频摘要方法有更好的性能指标。与表 1 中表现较好的 GANdpp 和 DR-DSN 以及 DTR-GAN 相比，虽然后者采用了 LSTMs 产生的网络对抗来进行视频摘要。但是文中的方法和其相比依然不差。对于 DR-DSN 来说，在 SumMe 和 TVSum 数据集上，F-score分别提高了 4%和 3%左右。与 DTR-GAN 相比也有 1.4%和 1.6%的提高。在和新近的 Cycle-SUM 和 Reg 比较时文中的方法依然有不错的表现。SUM-GAN-AAE 因为有注意力机制的加入在 SumMe 数据集上表现要比文中的方法好，但是在 TVSum 数据集上，文中的方法在 F-score 依然有 2%左右的提升。这充分说明基于差异特征的强化学习视频摘要方法更能全面有效的提取和利用原视频的信息。

表 1　F-score 实验结果对比

Tab. 1　Comparison of F-score experiment results

方法	SumMe	TVSum
sampling	33.4	15.5
Vsumm	33.7	/
GANdpp	39.1	51.7
DR-DSN	41.4	57.6
Reg	40.1	56.3
Cycle-SUM	41.9	57.6
DTR-GRN	44.6	59.1
SUM-GAN-AAE	48.9	58.3
单路原始特征	44.5	58.9
本文方法	46	60.7

3　结束语

本文提出了一种基于差异特征的强化学习视频摘要的方法，阐述如何有效利用提取视频帧的图像特征，建立视频帧对象之间的联系，对视频帧包含的信息达到一个长期有效的记忆的方式。通过联合相邻帧间的差异信息来有效地进行关键帧提取，达到预期的实验效果。由实验结果可以看出，本文提出的方法在两个标准数据总体性能表现优越。

参考文献

[1] TRUONG B T, VENKATESH S. Video abstraction: a systematic review and classification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2007, 3(1):1-37.

[2] LI L D, ZHOU K, XUE G R, et al. Video summarization via transferrable structured learning [C]//The 20th International Conference on World Wide Web, 2011: 287-296.

[3] SONG Y, VALLMITJNA J, STENT A, et al. TVSUM: summarizing web videos using titles[C]//The IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5179-5187.

[4] LI S W, PURUSHOTHAM S, CHEN C, et al. Measuring and predicting tag importance for image retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence , 2017, 39(12):2423-2436.

[5] HUSSAIN T, MUHANNAD K, UIIAH A, et al. Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM[J]. IEEETransaction on Industrial Information, 2019, 16(1):77-86.

[6] ZHONG S H, WU J X, JIANG J M. Video summarization via spatio-temporal deep architecture[J]. Neurocomputing, 2019, 332(1):224-235.

[7] SAHU A, CHOWDHURY A S. Multiscale summarization and action ranking in egocentric videos[J]. Patter Recognition Letter, 2020, 133(1):256-263.

[8] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning[J]. Machine Learning, 1992, 8(3/4):229-256.