智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

结合多尺度注意力机制和双向门控循环网络的视频摘要模型

闫河, 刘灵坤, 黄俊滨, 张烨, 段思宇

引用本文:
闫河,刘灵坤,黄俊滨,张烨,段思宇. 结合多尺度注意力机制和双向门控循环网络的视频摘要模型[J]. 智能系统学报, 2024, 19(2): 446–454.
YAN He, LIU Lingkun, HUANG Junbin, et al. Video summarization model based on the multiscale attention mechanism and bidirectional gated recurrent network[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 446–454.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202209048

DOI: 10.11992/tis.202209048
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20231113.1426.005

结合多尺度注意力机制和双向门控循环网络的视频摘要模型

闫河，刘灵坤，黄俊滨，张烨，段思宇
（重庆理工大学 两江人工智能学院，重庆 401135）

摘　要：针对视频摘要任务中全局注意力在长距离视频序列上注意力值分布的方差较大，生成关键帧的重要性分数偏差较大，且时间序列节点边界值缺乏长程依赖导致的片段语义连贯性较差等问题，通过改进注意力模块，采用分段局部自注意力和全局自注意力机制相结合来获取局部和全局视频序列关键特征，降低注意力值的方差。同时通过并行地引入双向门控循环网络（bidirectional recurrent neural network, BiGRU），二者的输出分别输入到改进的分类回归模块后再将结果进行加性融合，最后利用非极大值抑制（non-maximum suppression, NMS）和核时序分割方法（kernel temporal segmentation, KTS）筛选片段并分割为高质量代表性镜头，通过背包组合优化算法生成最终摘要，从而提出一种结合多尺度注意力机制和双向门控循环网络的视频摘要模型 (local and global attentions combine with the BiGRU, LG-RU)。该模型在 TvSum 和 SumMe 的标准和增强数据集上进行了对比试验，结果表明该模型取得了更高的 F-score，证实了该视频摘要模型保持高准确率的同时可鲁棒地对视频完成摘要。

关键词：视频摘要；自注意力机制；重要性分数；长程依赖；计算机视觉；双向门控循环神经网络；非极大值抑制；核时序分割方法
中图分类号：TP391.41　　**文献标志码**：A　　**文章编号**：1673−4785(2024)02−0446−09

中文引用格式：闫河，刘灵坤，黄俊滨，等．结合多尺度注意力机制和双向门控循环网络的视频摘要模型 [J]．智能系统学报，2024, 19(2): 446–454.
英文引用格式：YAN He, LIU Lingkun, HUANG Junbin, et al. Video summarization model based on the multiscale attention mechanism and bidirectional gated recurrent network[J]. CAAI transactions on intelligent systems, 2024, 19(2): 446–454.

Video summarization model based on the multiscale attention mechanism and bidirectional gated recurrent network

YAN He, LIU Lingkun, HUANG Junbin, ZHANG Ye, DUAN Siyu
(Liangjiang College of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China)

Abstract: In the video summary task, the variance of global attention value distribution on long distance video sequences is large, the importance score of generating key frames is large, and the semantic coherence of fragments is poor due to the lack of long-range dependence on the boundary values of time series nodes. Herein, by improving the attention module, segmented local self-attention and global self-attention mechanisms are merged to acquire the key features of local and global video sequences and lower the variance of attention values. Concurrently, the bidirectional gated recurrent neural network (BiGRU) is introduced in parallel, the output is input into the enhanced classification regression module, and afterward, the results are additively fused. Lastly, nonmaximum suppression and kernel temporal segmentation methods are applied to filter fragments and segment them into high-quality representative shots. The final summary is created by the knapsack combinatorial optimization algorithm. The video summary model LG-RU, which integrates the multiscale attention mechanism and BiGRU, is developed and compared with TvSum and SumMe's standard and enhanced data sets. It is demonstrated that the model has a higher F-score, which verifies that this model can complete the video summary robustly while preserving high accuracy.

Keywords: video summary; self-attention mechanism; importance score; long-range dependence; computer vision; BiGRU; nonmaximum suppression (NMS); kernel temporal segmentation (KTS)

收稿日期：2022−09−23.　**网络出版日期**：2023−11−14.
基金项目：国家重点研发计划“智能机器人”重点专项项目 (2018YFB1308602)；国家自然科学基金面上项目 (61173184)；重庆市自然科学基金项目 (cstc2018jcyjAX0694).
通信作者：闫河. E-mail: yanhe@cqut.edu.cn.

视频摘要是通过分析视频的结构和内容剔除存在的时空冗余，从原始视频中提取具有高代表性的片段帧[1]，并将帧拼接形成连贯视频摘要。

视频摘要按照学习方式主要分为两类: 无监督和有监督学习。

无监督的视频摘要方法大多是基于非深度学习的, 该类方法主要利用聚类或者字典学习的思想将视频摘要看作子集选择问题进行优化, 获取帧的重要性分数。文献 [2] 通过建模字典学习实现视频摘要, 随后又将其化为子集选择问题。但无监督的算法没有考虑视频帧间的时序信息, 其评价标准采用启发式的诸如代表性、稀疏性和差异性等, 评价性能较差[3-4]。虽然无监督和弱监督方法已经取得了显著的效果, 但其无法从手动创建的摘要中进行学习[5]。有监督的视频摘要算法主要是采用循环神经网络来建模视频序列信息, 文献 [6] 在视频摘要领域使用循环神经网络, 并使用行列式点过程作为补充以增强视频摘要的多样性, 但是该模型常常难以处理较长的视频序列, 部分信息会在长距离的传输过程中丢失。文献 [7] 使用长短期记忆网络 (long short-term memory, LSTM) 和强化学习实现视频摘要技术, 设计了奖励函数评估生成摘要的多样性和代表性, 但由于其对帧级特征关注方式与人类感知存在较大差异导致其对帧级分数评估不准确。随着注意力机制的提出, 文献 [8] 最先将注意力机制引入到视频摘要方法中, 并提出融合视觉和听觉的注意力模型。随后的大量工作都是在其基础上改进的, 文献 [9] 提出基于视觉注意力的自适应关键帧提取模型, 模拟人的高层感知提取底层特征, 将光流计算出的运动信息作为动态注意力, 定义感兴趣的物体为静态注意力。

文献 [10] 发现传统光流进行显著性挖掘计算复杂度高, 遂采用时序的梯度来建模动态显著性。但早期的注意力模型[6-11] 获取的都是视频底层特征, 易受噪声干扰, 且大都需要计算光流, 计算复杂度较高[12]。随着深度学习的发展, 基于注意力机制的视频摘要也愈发成熟, 通过将注意力机制和其他领域模块结合进行了许多尝试。文献 [13] 通过融合单向 LSTM 和注意力机制提出了新的编解码器网络, 较好地解决了长序列视频信息丢失问题; 文献 [14] 提出了一种新的视频摘要模型,使用自注意力机制来建模不同视频帧之间的关系与以前的模型相比, 其更加简单并且容易并行化, 使用可学习的自我注意机制来模拟帧的依赖关系; 文献 [12, 15-16] 将自注意力与强化学习结合, 通过自注意力机制建模视频帧的重要程度, 提高了模型的学习效率; 文献 [17] 将注意力机制和随机森林回归结合, 加权融合二者的损失从而提高了摘要结果的准确性, 但容易过拟合; 文献 [18] 提出了利用 IndRNN(independently recurrent neural network) 和单层注意力机制分别作为宽度组件和深度组件, 加权融合了低级特征和时间依赖。以上方法[12-18] 对比于过往模型虽然有了较大提升并取得了不错的效果, 但都是采用的单向独立循环神经网络以及单层的注意力机制, 一方面不能得到具有双向的长程依赖信息, 另一方面不能获得更细致的变换场景下的局部特征, 且在片段预测阶段使用的筛选镜头方法单一, 无法很好地预测镜头边界。文献 [19] 选择基于排序学习的方法, 将把视频摘要的提取等价为视频帧对视频内容表示的相关度排序问题, 但忽略了部分的上下文联系, 缺乏时序性。文献 [5] 在首次将目标检测中的锚框应用于视频摘要片段预测中, 在镜头选择阶段使用非极大值抑制模型帮助筛选关键镜头, 提出了一种新的视频摘要模型基于无锚框的视频摘要网络 (anchor free detect to summarize network, AF-DSNet)。但其在特征处理过程中的全局注意力机制在长距离的视频上性能下降, 注意力值的方差较大, 远距离视频帧之间的依赖性关注度较低[11], 且得到的特征序列的时序性仍然较差。

本研究针对全局自注意力机制导致的注意力权重分布的方差较大问题以及片段边界值仍缺乏时间长程依赖信息的问题, 提出了一种结合多层注意力机制和双向门控循环网络的视频摘要模型 (local and global attentions combine with the Bi-GRU, LG-RU), 该模型的核心在于改用多尺度注意力机制和 BiGRU 网络的特征进行加性融合, 并在镜头筛选过程中使用非极大值抑制算法过滤低质量片段。在多层注意力机制模块中, 通过融合局部、全局 2 种不同粒度的注意力以及残差输入得到探索具有视频局部和全局依赖关系的序列, 从而获得更加准确的帧重要性分数; 在 BiGRU 网络中得到视频的长程时序依赖关系。和传统的方法相比, 本模型采用 BiGRU 而非传统循环神经网络, 在保证精度的情况下优化了训练中反向传播的计算效率, 且体量更小。在特征处理阶段, 通过调整注意力模块和双向门控循环网络模块特征所占比例, 有效的加权结合了二者优点, 并通过视频重要性分数等信息筛选镜头, 得到了更具代表性和时序关系的摘要。

1　LG-RU 网络模型

本研究所提出的模型——LG-RU 模型的基

本结构如图 1 所示。特征提取阶段使用 GoogleNet 进行视频帧的特征提取；在多尺度注意力模块中，将特征序列输入到一个全局注意力模块中得到全局特征信息；同时将特征输入分割为等长的 N 段输入到 N 个局部多头注意力模块中，从而获得分段内的分段注意力权重；通过将分段注意力权重、全局注意力权重和残差输入进行加性融合得到视频的全局依赖关系；选用 BiGRU 网络用于对视频的时序关系进行建模。分别通过分类回归得到帧级相关信息，再将建模后的长程依赖信息与注意力模块输出信息进行加权融合，利用 NMS 算法[20] 筛选视频段以及用 KTS 算法[21] 分割视频段得到镜头，最后通过动态规划算法选取关键镜头形成摘要。

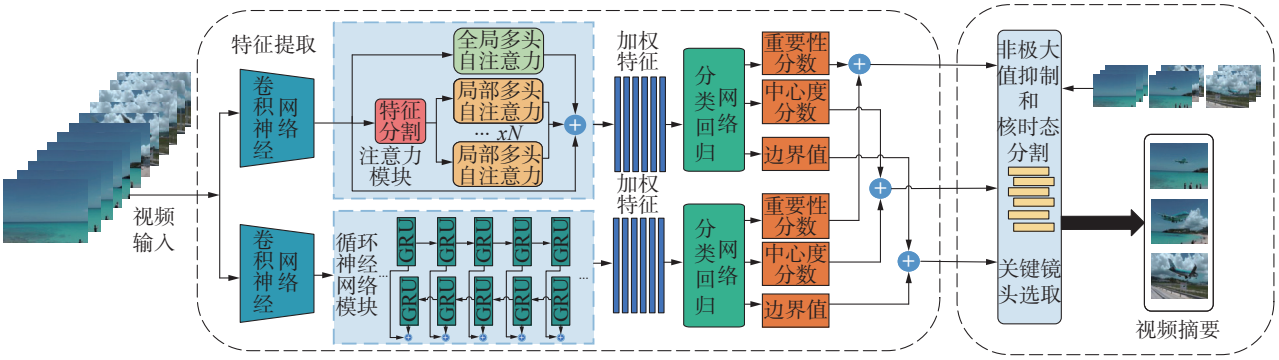


图 1　LG-RU 网络结构

Fig. 1　Main structure of the LG-RU

1.1　注意力模块

在输入视频序列 T_{in} 之后，使用预训练后的 GoogleNet 提取得到视频的特征序列 T。

由文献 [22] 工作可知，由于加性注意力机制没有考虑到输入序列的内部关系和点积注意力机制在输入向量维度较高时会有较高的方差，遂本研究采用缩放点积注意力机制来实现。在经过特征提取得到输入序列 T=(t_1, t_2,⋯, t_n) 后，在正向传播过程中根据输入序列生成 Query、Key、Value 3 个序列，首先计算序列中的元素查询 $\boldsymbol{q}_i$ 和每个键 $\boldsymbol{k}_i$ 的注意力得分 $\boldsymbol{e}_i$，计算公式如下

$$\boldsymbol{e}_i = \frac{\boldsymbol{q}_i^{\mathrm{T}} k_i}{\sqrt{d_k}} \tag{1}$$

其中，d_k 为序列维度数。再用 Softmax 函数对其进行归一化处理，得到 k_i 的权重 α_i

$$\boldsymbol{\alpha}_i = \text{Softmax}(\boldsymbol{e}_i) \tag{2}$$

最后将权重 α_i 和对应的值 v_i 加权求和得到注意力的输出，函数为

$$\text{Scaleattention}(Q,K,V) = \sum_i^N \boldsymbol{a}_i \boldsymbol{v}_i \tag{3}$$

本研究改进了 AF-DSNet 的原有方法，提出一种新的多尺度注意力机制模块 LG，在自注意力模块中，通过结合局部、全局注意力模块以及残差输入，特征融合方法选用加性融合，得到注意力模块的解码序列，如图 2 所示。在特征分割模块中，将序列 T 按固定长度 n 进行分段分割，得到分割段数 N

$$\begin{cases} N = 1,\ T \in (0,n) \\ N = T \div n + 1,\ T \in (n,+\infty),\ R \neq 0 \\ N = T \div n,\ T \in (n,+\infty),\ R = 0 \end{cases} \tag{4}$$

式中：

$$R = T\%n \tag{5}$$

其中，% 为取余计算。

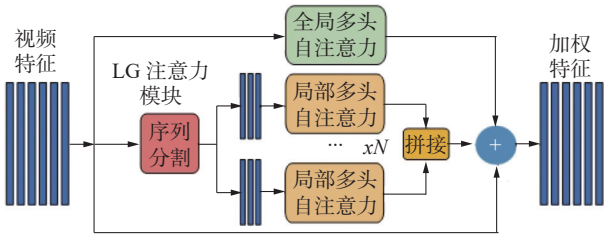


图 2　LG 注意力模块示意

Fig. 2　Diagram of LG attention module

通过对特征序列计算，得到局部注意力分段序列 $\boldsymbol{S}_i$，将 N 个局部分段 $\boldsymbol{S}_i$ 进行拼接得到局部序列 $\boldsymbol{S}$，其计算公式为

$$\boldsymbol{S} = \text{cat}(\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_N) \tag{6}$$

同样地得到全局注意力序列 W。根据本研究提出的序列计算方法为结合局部、全局注意力模块以及残差输入，注意力模块最终得到的视频序列 X 为

$$X = W + T + S \tag{7}$$

1.2　双向门控循环网络及分类回归模块

本研究在原 AF-DSNet 基础骨干之上增添了双向门控循环网络模块，用于建模序列的时序依赖。在双向门控循环网络模块中，选用 BiGRU 对输入序列 T=(t_1, t_2,⋯, t_n) 进行处理，将序列 T 同时

输入到正向与反向 GRU 中, 分别得到正反向状态序列 $\boldsymbol{Y}_i^a$ 和 $\boldsymbol{Y}_i^b$, 将 2 个状态序列连接得到每帧的前后向信息 $\boldsymbol{Y}_i=[\boldsymbol{Y}_i^a\ \ \boldsymbol{Y}_i^b]^{\mathrm{T}}$, 每帧的信息构成了输出序列 $\boldsymbol{Y}$=(Y_1, Y_2,⋯, Y_n)。

本研究改进 AF-DSNet 的分类回归网络提出了重要性边界网络 ICB-Net(important,central and boundary net, ICB-Net), 主要由共享的全连接层、ReLu、随机失活、层归一化以及 3 个并列的输出分支构成, 得到重要性分数是 S、中心度分数 V 和边界值 δ, 如图 3 所示。其中, 中心度分数 V 指的是预测帧与真实片段(ground truth)中心的偏移量, 偏移量越低说明预测帧的位置越准确, 其与重要性分数的乘积作为置信度分数可以作为衡量条件, 用于动态规划步骤中协助筛选关键镜头; 而边界值是通过监督学习视频帧的边界值二维向量来得到预测帧的左右边界值, 其表示了预测帧与真实片段边界的偏移量, 用于确定片段在时序序列线上的边界。此外, 以上得到的帧级相关信息将在非极大值模块筛选冗余片段时作为参数输入。

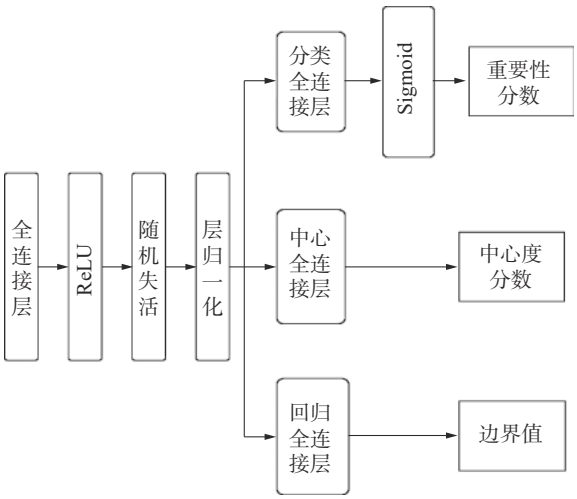


图 3　ICB-Net 结构

Fig. 3　ICB-Net structure

注意力模块和 BiGRU 模块结果序列分别输入到 ICB-NET 中, 得到 2 个模块的结果分别为重要性分数序列 $\boldsymbol{S}_a$、$\boldsymbol{S}_b$, 中心度分数序列 $\boldsymbol{V}_a$、$\boldsymbol{V}_b$, 边界值序列 $\boldsymbol{\delta}_a$、$\boldsymbol{\delta}_b$

$$\begin{aligned}\boldsymbol{S}_a,\boldsymbol{V}_a,\boldsymbol{\delta}_a&=\mathrm{ICBNET}(\boldsymbol{X})\\ \boldsymbol{S}_b,\boldsymbol{V}_b,\boldsymbol{\delta}_b&=\mathrm{ICBNET}(\boldsymbol{Y})\end{aligned}\tag{8}$$

式中, $\boldsymbol{S}_a$ 和 $\boldsymbol{S}_b$、$\boldsymbol{V}_a$ 和 $\boldsymbol{V}_b$、$\boldsymbol{\delta}_a$ 和 $\boldsymbol{\delta}_b$ 分别为注意力模块和双向门控循环网络模块各自的重要性分数、中心度分数以及边界值:

$$\begin{aligned}\boldsymbol{S}_i&=\alpha\boldsymbol{S}_a+\beta\boldsymbol{S}_b\\ \boldsymbol{V}_i&=\alpha\boldsymbol{V}_a+\beta\boldsymbol{V}_b\\ \boldsymbol{\delta}_i&=\alpha\boldsymbol{\delta}_a+\beta\boldsymbol{\delta}_b\\ \alpha+\beta&=1\end{aligned}\tag{9}$$

式中: α 和 β 为超参数, 训练过程中不断的手动调整 α 和 β 来优化模型, 在 α=0.4, 0.7 时在 SumMe 和 TvSum 上分别取得最优值, 后续试验部分对此进行了验证。

1.3　损失计算和关键镜头选择

1.3.1　损失计算

在损失函数的选择上, 通过计算均方误差损失 (mean square error loss) 和计算 tIoU 函数得到损失 L_1 和 L_2, 再利用二元交叉熵 (binary cross entropy, BCE) 损失 L_{center} 计算中心度损失 L_3, 其计算公式为

$$L_3=\frac{1}{N}\sum_e L_{\mathrm{center}}(V_e,V_e^*)\tag{10}$$

将中心度分数 V_e 和原始中心度分数 V_e^* 作为参数, 其中原始中心度分数计算为

$$V_e^*=\frac{\min(\delta_l^*,\delta_r^*)}{\max(\delta_l^*,\delta_r^*)}\tag{11}$$

式中: δ_l^*和δ_r^*分别代表第 e 个关键帧的左右边界。

并最终将 3 部分损失加权相加得到最终的损失量 L:

$$L=L_1+\lambda L_2+\gamma L_3\tag{12}$$

通过调整超参数 λ 和 γ 的值来获得最佳的效果。

1.3.2　关键镜头选择

本研究将阈值方法用于过滤视频, 通过阈值筛选去除视频冗余片段, 以期得到高质量的片段集合。该方法考虑重要性得分 $S=(S_1,S_2,\cdots,S_n)$及帧间相似度 $S_{\mathrm{im}}=(S_{\mathrm{im}_1},S_{\mathrm{im}_2},\cdots,S_{\mathrm{im}_{n-1}})$,其中 S_{im_i}的计算为

$$S_{\mathrm{im}_i}=\frac{X_i\cdot X_{i+1}}{\|X_i\|\times\|X_{i+1}\|}\tag{13}$$

过滤与已选视频帧相似度高于阈值的视频帧, 使得筛选出的视频片段集 C 具有高重要性得分的同时保持不同关键帧之间更低的相似度。

由 1.2 小节, 模块得到了序列的重要性分数 S、中心度分数 V 和边界值 δ 后, 再通过非极大值抑制过滤低质量冗余片段, 得到拥有起始时间片段的片段集 C:

$$C=\mathrm{filter}(S,V,\delta,S_{\mathrm{im}})\tag{14}$$

利用基于核的时序分割变化点检测模型 (kernel temporal segmentation, KTS) 对视频片段进行转换成为镜头, 来估计其重要性分数。镜头级重要性分数由该视频段内所有帧的重要性分数累加取平均, 其计算公式为

$$y_h=\frac{1}{n_h}\sum_{r=1}^{n_h}S_h^r\tag{15}$$

式中: n_h 是第 h 个镜头的长度; S_h^r是第 r 个视频帧的重要性分数。

$$\max\sum_{h=1}^{c}u_hy_h,\ \ \mathrm{s.t.}\sum_{h=1}^{c}u_hn_h\leqslant 15\%\times L_{\mathrm{en}}\tag{16}$$

式中：u_h 取 0 或 1，表示第 h 个视频镜头是否被选择；c 是镜头数量；L_{en} 是原始视频的长度，根据文献 [12] 的工作，生成摘要的长度限制为原始视频长度的 15%。随后通过动态规划模型选择关键镜头从而得到最终的视频摘要。

2　试验结果与分析

2.1　数据集与评价指标

2.1.1　数据集

本研究试验使用了 4 个公开的视频数据集，SumMe[23]、TvSum[24]、OVP[25] 和 YouTube[26]，数据集视频主要由节假日介绍、美食制作、体育运动等类型组成。其中后两者用于提高训练的数据集规模。帧级重要性分数为视频帧评分，数值为 0~1，数据集中的重要性分数由若干人打分加权求取。由于 OVP 和 YouTube 数据集的标注信息是关键帧，将关键帧设置为正类 1，非关键帧设置为负类 0 进行试验。标准数据集信息如表 1 所示。

表 1　标准数据集信息

Table 1　Standard data set information

数据集	视频数量	视频长度/min	标注信息
SumMe	25	1~6.5	重要性分数
TvSum	50	1~10	重要性分数
OVP	50	1~4	关键帧
YouTube	25	1~6.5	关键帧

2.1.2　评价指标

本研究遵循前人工作评估方法，采用 F-score 来评估模型的优劣。设 K 为模型生成的摘要，G 为人工标注的摘要，其重叠部分设为 O，如图 4 所示。

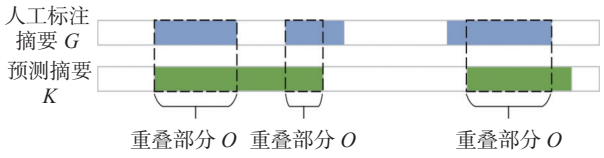


图 4　人工标注摘要和预测摘要重叠部分

Fig. 4　Overlap of manual annotated summary and predicted summary

准确率 (P_{recesion}) 和召回率 (R_{ecall}) 的计算公式如下

$$P_{\text{recesion}} = \frac{O}{S},\ R_{\text{ecall}} = \frac{O}{G} \tag{17}$$

根据准确率和召回率可以计算出 F-score (F_{score}) 的值。

$$F_{\text{score}} = 2 \times \frac{P_{\text{recesion}} \times R_{\text{ecall}}}{P_{\text{recesion}} + R_{\text{ecall}}} \tag{18}$$

2.2　对比试验与结果分析

为验证本研究提出的序列分割方法，将分割长度和融合方法作为变量在 2 个标准数据集上进行了试验，结果如表 2 和表 3 所示。

表 2　SumMe 数据集下不同分段长度下 F-score 值的大小

Table 2　F-Score values of different segment lengths in SumMe dataset　%

融合方法	分段长度/帧		
	100	200	400
加性融合	49.87	**51.84**	48.36
乘性融合	48.61	49.31	46.37
最大池化	49.36	50.12	48.33
平均池化	50.42	50.85	49.13

注：黑体代表最优结果，下同。

表 3　TvSum 数据集下不同分段长度下 F-score 值的大小

Table 3　F-Score values of different segment lengths in TvSum dataset　%

融合方法	分段长度/帧		
	100	200	400
加性融合	60.92	**62.18**	59.44
乘性融合	58.79	59.12	56.52
最大池化	59.28	60.73	58.40
平均池化	60.38	61.52	57.95

通过试验得出，在特征分割模块中，由式 (4) 中分段函数计算方式，当视频分段长度 m 取值为 200 帧时，本研究提出模型的评估指标 F-score 取得最优；当 m 取其他值时，会导致分割视频分段数量过多或过少，影响最终摘要的生成。

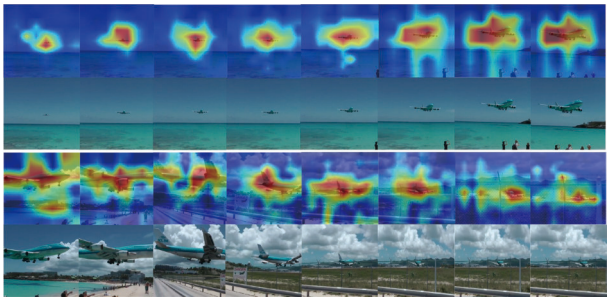
表 4 给出了最新的若干视频摘要模型的对比结果，数据均来源于原论文。在标准模式中，训练集、验证集、测试集都是来自同一种数据集；而在增强模式中，对于某种数据集，随机 20% 的数据用于测试，将该数据集剩下 80% 的数据和另外 3 种数据集共同构成训练集和验证集，相比于标准模式，该模式扩大了数据集的规模。根据表 4 数据，本模型 LG-RU 在 TvSum 和 SumMe 原始和增强数据集上均有较好效果。对比现有的视频摘要模型，LG-RU 具更好的表现。1）TvSum 数据集视频具有更长视频信息，更频繁的场景变换，LG-RU 中的 BiGRU 可以提取更多长距离上下文依赖关系；2）SumMe 数据集多为视频内容变化缓慢的单镜头原始视频，多层自注意力机制中不同粒度的自注意力模块可以很好地提取其内在关系和全局关系，减少了视频帧距离过长导致的注意力值方差过大问题；3）LG-RU 模型在镜头筛选中采用的非极大值抑制（non-maximum suppression，NMS）模型筛除了帧级分数差异较大的劣质片段。由图 5 可以看到，使用 LG 注意力模块前后的热力图变化，也反映了注意力分数的变化，使用 LG 模块之后的注意力分数评估的准确度有显著提升。是由于原有的注意力机制在对关键帧的

注意力分数打分的过程中，对于时间线上距离相隔较远的帧仍有权重分配，导致关键帧的分值受到较远帧的权重影响，增加了注意力分数的方差，降低了关键帧上注意力分数的准确性。而引入 LG 模块的动机是从序列中选取好的局部，并非仅为每个帧分配一个全局权重，而是从局部入手，局部分配权重再加权求和得到更准确的注意力分数，优化了关键帧注意力分数的打分过程，提高了重要性分数的评估准确度。

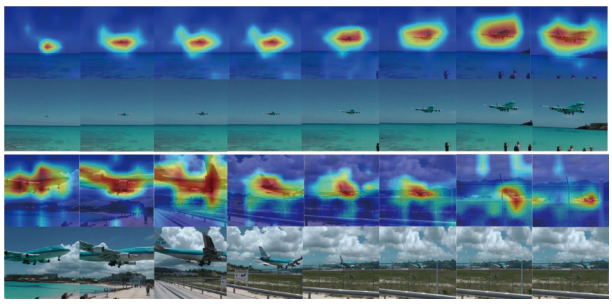
表 4　与当前先进摘要模型 F-score 对比

Table 4　Comparing with the F-score of the current advanced summary model　　%

模型名称	TvSum		SumMe	
	标准	增强	标准	增强
AF-DSNet[5]	61.9	62.2	51.2	53.3
vsLSTM[6]	54.2	57.9	37.6	41.6
dppLSTM[6]	54.7	59.6	38.6	42.9
DR-DSN[7]	58.1	59.8	42.1	43.9
A-AVS[11]	59.4	60.8	43.9	44.6
M-AVS[11]	61.0	61.8	44.4	46.4
VASNet[13]	61.4	62.4	49.7	51.1
WD-SN[18]	61.19	61.96	48.34	49.13
GMPAVS[26]	61.73	62.52	49.92	52.15
ALRSN[27]	61.86	63.06	50.71	52.61
LG-RU	**62.18**	**63.25**	**51.84**	**54.01**



(a) 使用 LG 注意力模块前

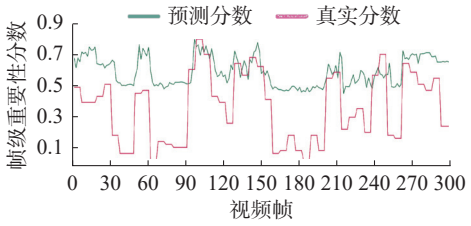


(b) 使用LG 注意力模块后

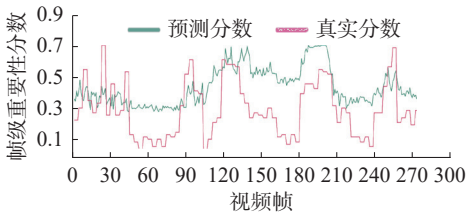
图 5　LG 模块使用前后 GradCAM 热力图

Fig. 5　LG module uses front and rear GradCAM heat maps

视频 20（video-20）的主要内容为炸鸡芝士汉堡的烹饪过程，视频 42（video-42）的主要内容为一对兄弟展示杂技摩托秀。图 6 中“真实分数”为人工标注摘要得分，“预测分数”为 LG-RU 模型预测得分，可以看出 2 条曲线之间的重要性分数趋势大致一致，并且模型预测曲线对于关键帧打分更高。这说明 LG-RU 模型可以很好地模仿人工标注方式，有效地识别关键镜头。



(a) video_20

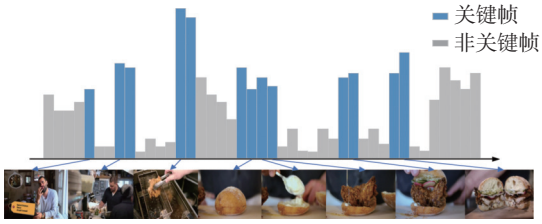


(b) video_42

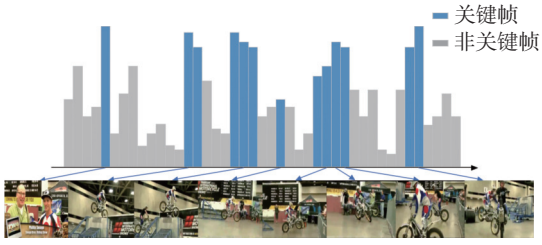
图 6　人工标注摘要和 LG-RU 预测摘要打分对比

Fig. 6　Comparison of scoring between manual annotated abstracts and LG-RU predicted abstracts

图 7 给出了模型视频摘要的结果，图中柱状条为人工标注摘要重要性分数，蓝色柱状条为通过 LG-RU 模型后选择出的关键镜头，以上镜头基本包含了活动事件的开头、高潮和结尾部分，并且所选镜头分数基本是高重要性分数镜头。式 (9) 中 α 与 β 影响着权重特征的融合，进而影响视频帧重要性分数判定，影响摘要的生成。



(a) video_20



(b) video_42

图 7　LG-RU 模型对视频进行摘要的结果

Fig. 7　Results of video summarization by LG-RU model

图 8 为不同 α 与 β 下 LG-RU 模型在 TvSum 数据集和 SumMe 数据集上的 F-score 值变化图，

通过对比试验，在 α=0.4, β=0.6 时在 TvSum 数据集上取得最佳效果值 0.632 5；在 α=0.7, β=0.3 时在 SumMe 数据集上取得最佳效果值 0.540 1。

图 9 给出了式 (11) 中 λ 和 γ 在不同数据集上对 F-score 值的影响，通过试验数据表明，在 λ=1.2 且 γ=0.9 时取得最佳效果。

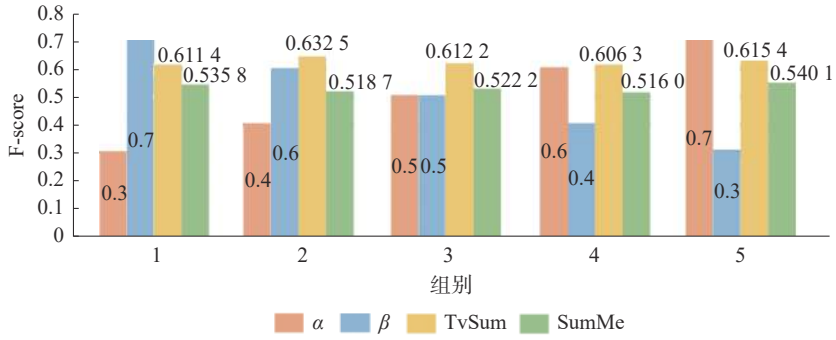
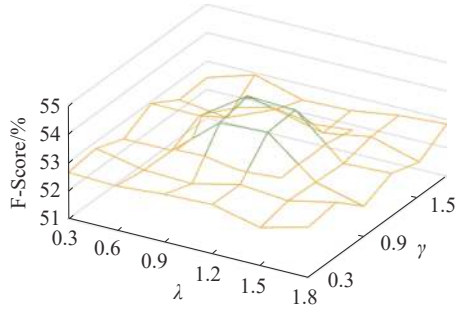
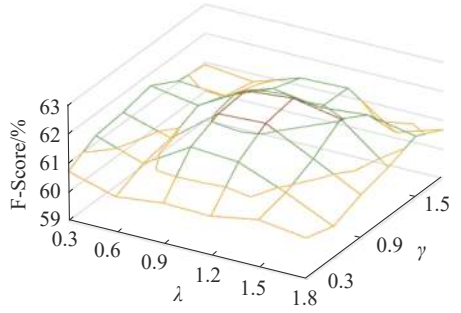


图 8　参数 α 与 β 在不同数据集上对 F-score 值的影响

Fig. 8　Influence of parameters α and β on F-score in different datasets



(a) SumMe(A) 数据集



(b) TvSum(A) 数据集

图 9　参数 λ 和 γ 在不同数据集上对 F-score 值的影响

Fig. 9　Influence of parameters λ and γ on F-score in different datasets

2.3　消融试验

表 5 给出了循环神经网络模块中分别选用 LSTM、BiLSTM 以及 BiGRU 网络的不同效果，可以看到，在增强数据集上，选取 BiGRU，在 TvSum 和 SumMe 数据集上都得到了更好的效果。在 BiGRU 取得更好效果的原因在于其双向机制以及轻量化结构，其取得的具有时序信息的特征序列还包含了前后向的语义指导信息，使得对于视频的理解更具有类人处理方式，通过前后向信息共同指导关键帧选择，从而影响了镜头选择，获得更加接近标注的结果。

为验证各模块的有效性，本研究基于 TvSum 和 SumMe 数据集进行消融试验，如表 6 所示。模型采用 AF-DSNet 作为骨干网络，超参数取最佳值，在保留其他代码细节不变的情况下，引入 LG 注意力模块、BiGRU 模块和 ICB-Net 模块。根据表中数据，本研究方法对比当前最先进的方法 AF-DSNet，在 2 个标准数据集上分别提高了 0.28% 和 0.64%，并在其 2 个增强数据集上提高了 1.05% 和 0.71%，结果表明本模型取得了更高的 F-score 分数，证实了本模型保持高准确率的同时可鲁棒地对视频完成摘要。

表 5　不同循环神经网络选择的 F-score 值

Table 5　F-score values selected by different recurrent neural networks　%

网络选择	TvSum	SumMe
LSTM	61.54	52.08
BiLSTM	62.08	53.74
BiGRU	63.25	54.01

表 6　不同模块在 TvSum 和 SumMe 数据集上消融试验的 F-score 值

Table 6　F-score values for ablation experiments of different modules on the TvSum and SumMe datasets　%

模块名称			TvSum		SumMe	
LG	BiGRU	ICB-Net	标准	增强	标准	增强
—	—	√	61.98	62.13	49.25	53.49
—	√	—	61.83	62.07	49.77	53.32
√	—	—	61.96	62.17	49.18	53.34
√	√	—	62.11	62.33	50.89	53.76
√	—	√	62.08	62.31	50.73	53.51
—	√	√	61.98	62.28	50.81	53.63
—	—	—	61.9	62.2	51.2	53.3
√	√	√	**62.18**	**63.25**	**51.84**	**54.01**

3　结束语

对于视频摘要生成任务, 本研究提出了一个多层自注意力机制和双向门控循环网络结合的视频摘要模型, 通过多层自注意力机制获取序列的全局和局部特征信息, 加权融合经过双向门控循环网络得到的具有长程时间依赖的信息, 有效地结合两部分模型的长处, 利用非极大值抑制算法过滤片段, 最终提高模型选择镜头的准确性。试验证明了该方法的有效性和可行性, 但该方法仅在常规的数据集下训练验证, 期望能在未来工作中拓展范围, 扩大影响。

参考文献:

[1] 冀中, 江俊杰. 基于解码器注意力机制的视频摘要 [J]. 天津大学学报(自然科学与工程技术版), 2018, 51(10): 1023–1030.
JI Zhong, JIANG Junjie. Video summarization based on decoder attention mechanism[J]. Journal of Tianjin University (science and technology edition), 2018, 51(10): 1023–1030.

[2] ELHAMIFAR E, VIDAL R. Sparse subspace clustering: algorithm, theory, and applications[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(11): 2765–2781.

[3] ELHAMIFAR E, SAPIRO G, SASTRY S S. Dissimilarity-based sparse subset selection[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(11): 2182–2197.

[4] ELHAMIFAR E, DE PAOLIS KALUZA M C. Subset selection and summarization in sequential data[M]. [S.l.]: Advances in Neural Information Processing Systems, 2017, 30.

[5] ZHU Wencheng, LU Jiwen, LI Jiahao, et al. DSNet: a flexible detect-to-summarize network for video summarization[J]. IEEE transactions on image processing, 2021, 30: 948–962.

[6] ZHANG Ke, CHAO Weilun, SHA Fei, et al. Video summarization with long short-term memory[M]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 766-782.

[7] ZHOU Kaiyang, QIAO Yu, XIANG Tao. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 7582-7589.

[8] MA Yufei, LU Lie, ZHANG Hongjiang, et al. A user attention model for video summarization[C]//Proceedings of the tenth ACM International Conference on Multimedia. New York: ACM, 2002: 533-542.

[9] JIANG Peng, QIN Xiaolin. Keyframe-based video summary using visual attention clues[J]. IEEE MultiMedia, 2010, 17(2): 64–73.

[10] EJAZ N, MEHMOOD I, BAIK S W. Efficient visual attention based framework for extracting key frames from videos[J]. Signal processing:image communication, 2013, 28(1): 34–44.

[11] ZHU Wencheng, LU Jiwen, HAN Yucheng, et al. Learning multiscale hierarchical attention for video summarization[J]. Pattern recognition, 2022, 122: 108312.

[12] 李依依, 王继龙. 自注意力机制的视频摘要模型 [J]. 计算机辅助设计与图形学学报, 2020, 32(4): 652–659.
LI Yiyi, WANG Jilong. Self-attention based video summarization[J]. Journal of computer-aided design & computer graphics, 2020, 32(4): 652–659.

[13] JI Zhong, XIONG Kailin, PANG Yanwei, et al. Video summarization with attention-based encoder–decoder networks[J]. IEEE transactions on circuits and systems for video technology, 2020, 30(6): 1709–1717.

[14] FAJTL J, SOKEH H S, ARGYRIOU V, et al. Summarizing videos with attention[M]//Computer Vision-ACCV 2018 Workshops. Cham: Springer International Publishing, 2019: 39-54.

[15] MAHASSENI B, LAM M, TODOROVIC S. Unsupervised video summarization with adversarial LSTM networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2982-2991.

[16] LEI Jie, LUAN Qiao, SONG Xinhui, et al. Action parsing-driven video summarization based on reinforcement learning[J]. IEEE transactions on circuits and systems for video technology, 2019, 29(7): 2126–2137.

[17] 李雷霆, 武光利, 郭振洲. 自注意力机制和随机森林回归的视频摘要生成 [J]. 计算机工程与应用, 2022, 58(4): 198–205.
LI Leiting, WU Guangli, GUO Zhenzhou. Video summarization generation based on self-attention mechanism and random forest regression[J]. Computer engineering and applications, 2022, 58(4): 198–205.

[18] ZHOU Juanping, LU Lu. Wide and deep learning for video summarization via attention mechanism and independently recurrent neural network[C]//2020 Data Compression Conference. Snowbird: IEEE, 2020: 407.

[19] 王鈅润, 聂秀山, 杨帆, 等. 基于排序学习的视频摘要 [J]. 智能系统学报, 2018, 13(6): 921–927.
WANG Xingrun, NIE Xiushan, YANG Fan, et al. Video summarization based on learning to rank[J]. CAAI transactions on intelligent systems, 2018, 13(6): 921–927.

[20] XU Huijuan, DAS A, SAENKO K. Two-stream region convolutional 3D network for temporal activity detection[J]. IEEE transactions on pattern analysis and machine intelli-

gence, 2019, 41(10): 2319–2332.

[21] POTAPOV D, DOUZE M, HARCHAOUI Z, et al. Category-specific video summarization[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 540−555.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] //Proceedings of the 31st International Conference on Advances in Neural Information Processing Systems. New York: Curran Associates, Inc, 2017: 5998-6008.

[23] GYGLI M, GRABNER H, RIEMENSCHNEIDER H, et al. Creating summaries from user videos[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 505−520.

[24] SONG Yale, VALLMITJANA J, STENT A, et al. TVSum: Summarizing web videos using titles[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 5179−5187.

[25] DE AVILA S E F, LOPES A P B, DA LUZ A, et al. VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method[J]. Pattern recognition letters, 2011, 32(1): 56–68.

[26] 王坤阳, 高伟, 滕国伟. 基于门控多头注意力机制的视频摘要 [J]. 工业控制计算机, 2022, 35(12): 120–122.
WANG Kunyang, GAO Wei, TENG Guowei. Video summarization based on gated multi-head attention mechanism[J]. Industrial control computer, 2022, 35(12): 120–122.

[27] 梅锋, 周娟平, 陆璐. 结合局部奖励机制的视频摘要技术研究 [J]. 计算机工程与应用, 2021, 57(11): 211–218.
MEI Feng, ZHOU Juanping, LU Lu. Research on video summarization technology combining local reward mechanism[J]. Computer engineering and applications, 2021, 57(11): 211–218.

作者简介：

闫河，博士，教授，主要研究方向为图像多尺度几何分析、目标跟踪、模式识别。主持国家自然科学基金面上项目、中国博士后基金项目各 1 项，重庆市自然科学基金项目、教育部重点实验室访问学者基金项目各 2 项；以单位负责人参加科技部“十三五”重点研发计划“智能机器人”重点专项项目 1 项；参研省部级项目 10 余项。发表学术论文 90 余篇。E-mail：yanhe@cqut.edu.cn。

刘灵坤，硕士研究生，主要研究方向为与深度学习相结合的视频摘要处理、视频理解、目标检测。E-mail：LiuLingK@stu.cqut.edu.cn。

黄骏滨，硕士研究生，主要研究方向为与深度学习相结合的视频摘要处理和视频描述方法。E-mail：huangjunbin@2020.cqut.edu.cn。