DOI:10.11784/tdxbz201801077

基于解码器注意力机制的视频摘要

冀　中，江俊杰
(天津大学电气自动化与信息工程学院，天津 300072)

摘　要：作为一种快速浏览和理解视频内容的方式，视频摘要技术引起了广泛的关注. 本文将视频摘要任务看作是序列到序列的预测问题，设计了一种新颖的基于解码器的视觉注意力机制，并基于此提出一种有监督视频摘要算法. 所提方法考虑到视频帧之间的内在关联性，利用长短时记忆网络将注意力集中在历史的解码序列，融合历史的解码信息有效地指导解码，提升模型预测的准确性. 所提算法主要在 TVSum 和 SumMe 数据集上进行了大量实验，验证了其有效性及先进性.

关键词：视频摘要；视觉注意力模型；编解码模型；长短时记忆网络
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：0493-2137(2018)10-1023-08

Video Summarization Based on Decoder Attention Mechanism

Ji Zhong，Jiang Junjie
(School of Electrical and Information Engineering，Tianjin University，Tianjin 300072，China)

Abstract：As a way to quickly browse and understand video content，video summarization has attracted wide attention. This paper treats video summarization as a sequence-to-sequence prediction problem and proposes a novel visual attention model based on decoder，which is further applied to supervised video summarization. The proposed method pays attention to decoding sequence by using long short-term memory network. It considers the intrinsic association between video frames，and utilizes the previous decoding sequences to effectively guide the decoding process，which improves the prediction accuracy. Extensive experiments are mainly conducted on TVSum and SumMe datasets，which demonstrate the effectiveness and superiority of the proposed method.

Keywords：video summarization；visual attention model；encoder-decoder model；long short-term memory network

近年来，如何对海量视频数据进行快速有效的浏览、检索和分析成为了多媒体分析领域的研究热点. 视频摘要是其中一项具有重要理论和实际应用价值的技术，受到了广泛的关注. 它是指利用智能分析技术分析视频结构、理解视频内容，并从原始的数据中选取具有代表性的、有意义的部分，将它们以某种方式组合并生成紧凑的、用户可读的原始视频的缩略[1]. 依据摘要的最终呈现形式，视频摘要通常分为两种形式：基于关键帧(key frames)的静态视频摘要和基于关键镜头(key shots)的动态视频摘要，本文关注的是后者.

目前视频摘要的研究大多基于无监督的学习方法，包括聚类[2-5]、图模型[6-9]、稀疏编码[10-12]等方法. 近几年，研究者开始聚焦于有监督的视频摘要方法[13-19]. 有监督的方法直接从人工标注的视频摘要学习选取摘要的准则，使摘要的自动生成方式类似于人选取摘要的决策过程，摘要的结果更接近人类的理解方式. 这类方法要解决的问题是如何从原始视频帧序列中提取关键帧或者关键镜头序列，其本质可视为一种序列到序列(sequence-to-sequence，Seq2Seq)

收稿日期：2018-01-22；**修回日期**：2018-03-13.
作者简介：冀　中(1979—　　)，男，博士，副教授.
通讯作者：冀　中，jizhong@tju.edu.cn.
基金项目：国家自然科学基金资助项目(61472273，61771329).
Supported by the National Natural Science Foundation of China(No. 61472273 and No. 61771329).

的结构化预测问题[20]. 编解码器(ecoder-decoder)框架就是解决此类问题的有效方式之一. 在该框架中,编码器将输入序列编码成固定长度的中间向量,然后解码器再将其解码成符合任务需求的输出序列. 编码器和解码器一般采用循环神经网络(recurrent neural network, RNN)或长短时记忆(long short-term memory, LSTM)网络,尤其是 LSTM 在建模长期依赖性问题上有着极大的优势,能够深层次地挖掘对解决任务有用的序列信息.

用户在选取视频摘要时存在一种视觉注意力机制,即越受人眼关注的镜头或者视频帧,被选入摘要的可能性越大. 现有的一些工作试图对注意力机制进行建模,以此作为选取摘要的依据. 例如,Ma 等[21]利用视频的运动、脸部、相机聚焦以及声音等信息,分别以线性和非线性的方法融合这些信息构建注意力模型,指导视频摘要的生成. Ejaz 等[22]基于图像的显著性检测方法和时间梯度分别对静态注意力和动态注意力进行建模,然后非线性地融合两种注意力模型以此生成摘要. 然而,上述工作仅仅利用了底层特征,很难对人类抽象的视觉注意力机制进行建模. 而且现有方法均是利用无监督的方法人为地构建注意力模型,具有一定的局限性,无法较好地学习用户选取摘要时的注意力机制.

而基于注意力机制的编解码器框架已经在机器翻译[23-24]、文本摘要[25]、图像描述[26]、视频描述[27-28]等任务中有突出的表现. 如 Bahdanau 等[23]在处理英语翻译为法语的任务时,编码器采用双向 LSTM 对英语单词序列进行编码,解码器在预测每个法语单词时,会以不同的注意力权重关注编码序列中不同位置的编码向量,从而提升了翻译的准确性. 在文献[27]中,编码器将原始视频编码成视频特征序列,解码器利用视频特征序列生成描述性的语句,解码器在每一时刻生成单词时会关注编码序列中不同位置的视频特征.

受有监督的视觉注意力模型在机器翻译[23-24]、视频描述[27-28]等领域的启发,本文设计了一种新颖的视觉注意力机制,与编解码框架结合起来,提出了基于解码器注意力机制的有监督视频摘要算法.

本文的创新点有 2 个. ①设计了一种新颖的基于解码器的视觉注意力机制. 考虑到视频帧之间的内在关联性,利用长短时记忆网络将注意力集中在历史的解码序列,融合历史的解码信息有效地指导当前的解码过程,提升模型预测的准确性. ②将所提注意力机制与编解码器框架相结合,提出一种新的有监督视频摘要方法 SUM-attDecoder,并在主流的数据集上验证了其有效性与先进性.

1 相关工作

依据生成摘要的过程中是否需要标注信息,视频摘要的研究可分为无监督和有监督两大类方法. 其中无监督的视频摘要研究较早,常用的方法包括聚类[2-5]、图模型[6-9]、稀疏编码[10-12]等. 例如,VSUMM[2]通过对视频的颜色特征进行 k 均值聚类,并且通过聚类中心生成视频摘要. 为了建模视频帧间的高阶信息,文献[8]提出一种基于超图模型的视频摘要方法,通过对视频帧构建超图模型,然后在此基础上进行主集聚类得到视频摘要. Panda 等[9]将视频摘要视为图聚类问题,运用骨架图和随机游走方法对该问题建模. Mei 等[10]提出了最小稀疏重构的方法,通过最小化原始视频帧与候选关键帧之间的重构误差的原则来选取摘要,最终误差最小的候选关键帧可作为摘要. 文献[11]通过包含帧内视角和帧间视角的相关性的目标函数学习联合嵌入空间,然后结合学习到的嵌入空间,采用稀疏表征选择的方法生成多视点视频摘要. Li 等[29]结合视觉、音频信息,在最大边界相关的思想下设计了迭代选择关键镜头的视频摘要算法. 最近,基于生成视频摘要的内容和原视频内容尽可能相近的原则,文献[30]应用对抗生成网络(GAN)生成视频摘要,也取得了较好的性能.

有监督的视频摘要是基于原始视频的人工标注学习一个摘要选择器,使其提取的摘要最大程度地接近人工的选取标准. 例如,Gong 等[13]提出序列行列式点过程(sequential determinantal point process, seqDPP)方法,目的是最大程度地减少提取关键帧的冗余性,使提取的关键帧和人工摘要更接近. Zhang 等[14]重点考虑了相似视频的结构相关性,利用非参数方法学习从标注视频到测试视频的迁移摘要结构,以此来指导摘要的生成. Gygli 等[15]通过设计多目标函数,使生成的摘要能满足兴趣度、代表性、均匀性的评价标准. Li 等[16]设计了 4 个评价标准,分别是代表性、多样性、故事性和重要性,通过建立一个评分函数来线性地组合这 4 个评价指标作为生成摘要的指导准则. Potapov 等[18]首先将特定主题的目标视频分割成语义一致的视频片段,再用 SVM 分类器预测每个片段的分数,选择分数最高的视频片段作为摘要. Gygli 等[31]将视频先分割成超帧,再结合底层特征和高层特征来训练线性回归模型预测超帧的兴趣度,最后通过求解最大化超帧的兴趣度问题来选取视频摘要.

近年来，基于深度学习方法的视频摘要也引起了研究者的关注. 例如，Zhang 等[17]首次将视频摘要看作序列到序列问题，引入长短时记忆模型 LSTM 和多层感知器对视频帧序列进行建模，并设计了交叉熵和行列式点过程（determinantal point process，DPP）2 个目标函数分别保证选取摘要的重要性和多样性. Yang 等[32]结合自编码器和 LSTM 模型对视频序列进行建模，并设计了指数衰减的损失函数，用于提取视频中的精彩片段. 与已有方法不同，本文在 LSTM 模型的基础上引入了视觉注意力机制，在预测视频帧分数时充分利用历史解码信息，从而提升了模型预测的准确性.

2　所提方法

本文的算法主要包括编解码模型和关键镜头选取模型，如图 1 所示. 编解码模型由编码器和解码器构成，作用是将视频序列映射成重要性分数序列，预测每一视频帧的重要性程度. 关键镜头选取模型则根据视频帧的重要性分数，利用动态规划方法选取最优的视频镜头集合，生成视频摘要.

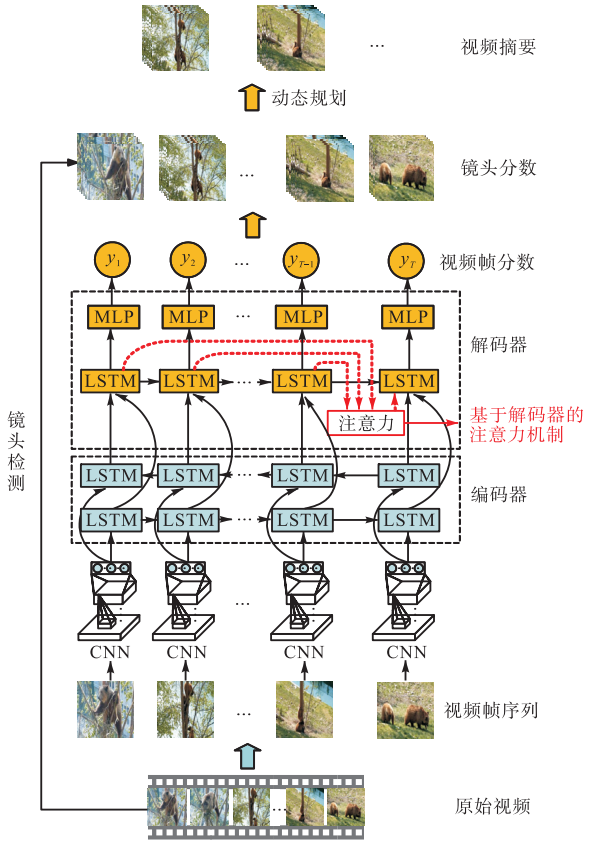

图 1　本文算法的示意

Fig.1　Architecture overview of the proposed method

2.1　编码器模型

本文将有监督的视频摘要任务看作是 Seq2Seq 的预测问题，输入序列是一段视频帧序列，输出序列是对应视频帧的重要性分数. 在解决 Seq2Seq 问题时，通常先将输入序列通过编码器转换成中间码的形式再映射成输出序列. 在有监督的视频摘要中，编码器的作用是将提取好的视频特征全部编码成能描述视频时序特性的向量序列，再将其送入解码器. 编码器可以表示为

$$\boldsymbol{v}=\phi_{W_{\text{enc}}}(X) \tag{1}$$

式中：$\phi_{W_{\text{enc}}}$ 代表以 W_{enc} 为参数的神经网络；v 是需要输入到解码器的中间码. 编码器可以采用任何神经网络，这取决于具体的任务以及输入数据. 在视频摘要任务中，考虑到人类在决策视频帧或者视频镜头能否被选入视频摘要时，都会依据它们的上下文信息来判断其重要性，即某一视频帧的重要性不仅受到前面时刻视频帧的影响，还受到后续时刻的视频帧的影响. 因此本文采用双向 LSTM（bidirectional LSTM，BiLSTM）网络作为编码器，能有效地捕捉视频序列的上下文信息，更有利于后续的解码. BiLSTM 包括正向 LSTM 和反向 LSTM，是对单层 LSTM 网络的进一步扩展，正向 LSTM 用来捕获前一时刻的隐藏信息，而反向 LSTM 用来捕获后一时刻的信息，两部分的信息拼接在一起作为后续解码器的输入.

编码器的输入序列是视频特征序列 $\boldsymbol{x}=(x_1,x_2,\cdots,x_T)$，该序列是由 T 个视频帧的特征组合而成. 编码器的正向 LSTM 首先正向输入序列 $\boldsymbol{x}$，从 x_1 到 x_T 逐步输入，输出正向的隐层状态 $(\boldsymbol{h}_1^{\text{v}},\cdots,\boldsymbol{h}_T^{\text{v}})$；与此同时，$\boldsymbol{x}$ 反向输入到反向 LSTM 中，输出反向的隐层状态 $(\boldsymbol{h}_1^{\text{w}},\cdots,\boldsymbol{h}_T^{\text{w}})$，最后将这两个隐层状态序列组合成编码序列 $\boldsymbol{v}_t=[\boldsymbol{h}_t^{\text{v}},\boldsymbol{h}_t^{\text{w}}]^{\text{T}}$，$\boldsymbol{v}_t$ 包含了输入的第 t 帧的前向信息和后向信息，所有输入序列的前后向信息构成编码序列 $\boldsymbol{v}=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_T)$，$\boldsymbol{v}$ 将作为后续解码器的输入.

2.2　所提解码器模型

解码器的作用是将编码器转换成的中间码映射成输出序列. 如果输出序列的长度大于 1，解码器一般要设计成可循环的结构，因为解码器需要通过获取已被预测出的历史信息来防止重复预测. 尤其是在引入注意力机制的模型中，历史的解码信息对当前解码过程的影响程度不同，而且不同时刻的注意力权重会有所不同. 因此，解码器必须能够储存历史信息.

本文采用的解码器是由 1 层 LSTM 网络和 2 个

全连接层构成，定义解码器的输出公式为

$$\boldsymbol{y}_i = g(\boldsymbol{s}_i) \tag{2}$$

式中：$\boldsymbol{s}_i$ 和 $\boldsymbol{y}_i$ 分别为解码器 i 时刻的隐藏状态(LSTM 网络的输出)和解码器的输出($\boldsymbol{y}_i$ 为一个数值，代表对应输入视频帧的重要性分数)；$g(\bullet)$ 函数代表两层的多层感知器，将 LSTM 网络的输出向量映射成数值，代表重要性分数. 本文设计了 2 种 s_i 实现方法，将在下文分别进行介绍.

2.2.1 SUM-LSTM 算法

为了验证视觉注意力机制对编解码器框架性能的提升，本文首先设计了一个基准模型，称为 SUM-LSTM，它由编码器和解码器构成，编码器的结构同第 2.2 节所述，解码器为未引入视觉注意力机制的单层 LSTM 网络和两个全连接层，结构如图 2 所示，该模型的 LSTM 网络公式为

$$\boldsymbol{s}_i = f(\boldsymbol{s}_{i-1}, \boldsymbol{v}_i) \tag{3}$$

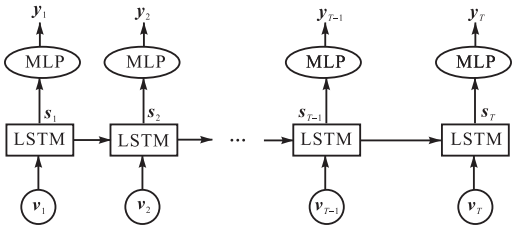

图 2 SUM-LSTM 解码器的结构

Fig.2 Illustration of the decoder structure in SUM-LSTM

2.2.2 SUM-attDecoder 算法

SUM-attDecoder 称为基于解码器视觉注意力机制的视频摘要算法，是在 SUM-LSTM 的基础上，在解码器部分引入了视觉注意力机制，编码器仍采用双向 LSTM 网络. 其结构如图 3 所示，其中解码器的 LSTM 网络 $\boldsymbol{s}_i$ 的公式定义为

$$\boldsymbol{s}_i = f(\boldsymbol{c}_i, \boldsymbol{s}_{i-1}, \boldsymbol{v}_i) \tag{4}$$

式中：$\boldsymbol{s}_{i-1}$ 既是 $i-1$ 时刻 LSTM 网络的隐藏状态，又是该网络 $i-1$ 时刻的输出；$\boldsymbol{v}_i$ 为 i 时刻解码器的输入；$\boldsymbol{c}_i$ 是上下文向量，由 i 时刻的注意力模型得到，是由 LSTM 在 $1,2,\cdots,i-1$ 时刻的输出向量加权融合而成的，即由 $\{\boldsymbol{s}_1,\cdots,\boldsymbol{s}_{i-1}\}$ 加权和得到. $\boldsymbol{c}_i$ 融合了之前所有时刻的历史输出信息，但是对每一历史时刻的注意力权重不同，视觉注意力机制会指导网络去学习不同时刻的注意力权重.

上下文向量 c_i 计算公式为

$$\begin{cases} \boldsymbol{c}_i = \sum_{j=1}^{i-1} \alpha_{ij} \boldsymbol{s}_j \\ \text{s.t.} \sum_{j=1}^{i-1} \alpha_{ij} = 1 \end{cases} \tag{5}$$

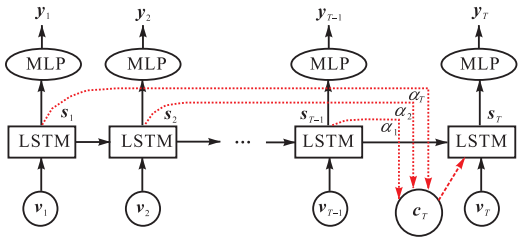

图 3 所提 SUM-attDecoder 解码器的结构

Fig.3 Illustration of the decoder structure in the proposed SUM-attDecoder

式中：$\boldsymbol{s}_j$ 为 LSTM 在 j 时刻的输出，即历史的解码信息，$j \in \{1,2,\cdots,i-1\}$；α_{ij} 表示 i 时刻的解码对 j 时刻历史解码序列 s_j 分配的注意力权值，即

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1}^{i-1} \exp(e_{ij})} \tag{6}$$

$$e_{ij} = a(\boldsymbol{s}_{i-1}, \boldsymbol{s}_j) \tag{7}$$

生成 e_{ij} 的函数是一个多层感知器，该网络的输入由 LSTM $i-1$ 时刻的隐藏状态 $\boldsymbol{s}_{i-1}$ 和 j 时刻的输出 $\boldsymbol{s}_j$ 构成，e_{ij} 代表 j 时刻解码的输出对 i 时刻解码过程的影响程度.

因为多层感知器的输出是数值，而参与 LSTM 网络内部循环运算的是向量，因此本文视觉注意力机制利用的历史解码信息指的是解码器中 LSTM 网络的历史输出向量，以便于注意力信息参与运算.

2.3 关键镜头选取模型

视频摘要的最终呈现形式是镜头集合，因此首先对视频进行镜头检测，分割成若干个视频镜头作为后续摘要提取的候选镜头. 本文采取基于核的时域分割(kernel temporal segmentation，KTS)[18] 镜头检测算法. 镜头检测后，利用所提 SUM-attDecoder 模型预测得到的视频帧重要性分数计算每个镜头的重要性分数，即对镜头包含视频帧的重要性分数求和.

与文献[17]相同，本文将视频摘要的长度限定为原视频长度的 15%，利用 0/1 背包算法求解最大化镜头集合的重要性分数的最优化问题，生成视频摘要. 具体地，给定所有镜头的集合 S，从中选取长度不大于 L_s 的子集，使得子集的重要性分数和最大，求解最优化问题

$$\begin{cases} \max\limits_{x} \sum_{i=1}^{n} x_i I(S_i) \\ \text{s.t.} \sum_{i=1}^{n} x_i |S_i| \leqslant L_s \end{cases} \tag{8}$$

式中：$x_i \in \{0,1\}$，$x_i = 1$ 代表第 i 个镜头被选入摘要

中；$I(S_i)$ 代表第 i 个镜头 S_i 的重要性分数.

3　实验结果及分析

3.1　实验数据

本文的实验用到了 4 个公开的标准数据集，分别是 SumMe[31]、TVSum[33]、YouTube[2]和 OVP[2]. 其中，后 2 个数据集主要用于数据增强实验的验证. SumMe 数据集包含 25 个视频，记录了节日、运动、重大事件等主题的内容，视频时长为 1 ~ 6 min. 这些视频都是用户拍摄的原始视频，并未做后期处理，包含较多的冗余信息. TVSum 数据集是通过在 YouTube 网站上搜索 10 个视频类别的关键词搜集的，它一共有 50 个视频，每类视频有 5 个，主题涵盖新闻、记录片、用户视频等，视频时长为 2 ~ 10 min. YouTube 数据集也是在 YouTube 网站上搜集的 1 ~ 10 min 视频，视频类型主要有动画片、新闻、运动、商业、电视节目和家庭视频，一共有 50 个视频. 但是考虑到动画片与其他类型的视频在内容、颜色、时长方面差异较大，不利于模型的训练和测试，所以与文献[17]相同，本文采用的 YouTube 数据集剔除了 11 个动画视频，保留剩下的 39 个视频作为数据集. OVP 数据集的 50 个视频均来自于 open video project 网站.

表 1 是对 4 个数据集的详细描述，其中 SumMe、TVSum 的标签是人工标注的重要性分数，而 YouTube、OVP 的标签是人工选取的关键帧，本文将其转化为重要性分数——对应关键帧的位置分数为 1，否则为 0. 数据集划分为训练集、验证集以及测试集，比例分别为 60%、20%、20%.

表 1　数据集的详细描述

Tab.1　Detailed description of dataset

数据集	视频数量	数据集描述	时长/min	标注
SumMe	25	用户拍摄视频	1 ~ 6	视频帧的重要性分数
TVSum	50	剪辑视频	2 ~ 10	视频帧的重要性分数
YouTube	39	网络视频	1 ~ 10	关键帧
OVP	50	记录片	1 ~ 4	关键帧

3.2　实验参数和评价指标

在预处理中对原始视频进行下采样，采样率为 2 帧/s. 为了便于与 vsLSTM[17]算法比较，笔者利用在 ImageNet 数据集上预训练的 GoogLeNet 网络提取特征，将该网络倒数第 2 层的输出向量作为视频帧的特征. 本文算法的 3 个 LSTM 层，每层都含有 256 个单元数，且多层感知器第 1 层含有 256 个单元，第 2 层含有 1 个单元. 训练时采用的最优化算法是随机梯度下降法，BatchSize 的大小为 16，视频帧样本序列的长度为 10，学习率为 0.001 5，本算法的目标函数是均方差函数. 对于 SumMe 和 TVSum 数据集，每个数据集上做 10 次实验，取 F 值的平均值作为该数据集的评价指标.

与文献[17]相同，采用 F 值(F-score)对结果进行评价，它可由精度(Pre)、召回率(Rec)计算得到，即

$$\mathrm{Pre}=\frac{N_{\mathrm{matched}}}{N_{\mathrm{AS}}} \tag{9}$$

$$\mathrm{Rec}=\frac{N_{\mathrm{matched}}}{N_{\mathrm{US}}} \tag{10}$$

$$\text{F-score}=\frac{2\mathrm{Pre}\cdot\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{11}$$

式中：N_{matched} 表示生成摘要与用户摘要匹配的长度，即生成的摘要中与用户摘要中重叠的视频帧数量；N_{AS} 表示自动生成摘要的长度，即摘要中包含视频帧的个数；N_{US} 表示用户摘要的长度. 精度反映了算法对自动生成的摘要是否匹配用户摘要的判断准确性，召回率反映了自动摘要与用户摘要匹配的能力，即用户摘要的覆盖率. F 值平衡了精度和召回率 2 个指标，是对视频摘要质量的一个整体评价指标.

3.3　实验分析与讨论

表 2 给出了所提算法与对比算法在 SumMe 和 TVSum 数据集上的实验结果比较，对比算法都是近几年提出的先进的无监督和有监督的视频摘要算法. 其中 vsLSTM[17]与本文 SUM-LSTM 方法都是基于编解码器框架，不同之处在于其解码器是全连接层，没有利用视觉注意力机制.

从表 2 中可以看出，本文所提算法在 2 个数据集上均取得了较高的性能. 在 SumMe 数据集上，所提 SUM-attDecoder 算法取得了最好的性能，比最好的对比算法 vsLSTM 高 0.6%. 而在 TVSum 数据集上，所提算法取得了次优的性能，比性能最好的 vsLSTM 低 1.3%. 由此可以看出，所提 SUM-attDecoder 算法生成的视频摘要质量是比较好的. 所提算法较 vsLSTM 具有更高的模型复杂度，需要更多类别的视频来训练网络，TVSum 数据集仅有 10 类视频，导致

训练模型的泛化性较差,因此所提算法在 TVSum 数据集上的性能低于 vsLSTM. 这也是表 3 中在数据增强后所提算法的性能反而超过 vsLSTM 的原因.

表 2 不同视频摘要算法的性能对比

Tab.2 Performance comparison between different video summarization methods

数据集	算法	有无监督	F 值
SumMe	MMR[29]	无	26.0
	VSUMM[2]	无	33.7
	vsLSTM[17]	有	37.6
	SUM-LSTM(本文)	有	35.1
	SUM-attDecoder(本文)	有	38.2
TVSum	TVSum[33]	无	50.0
	Zhao 等[34]	无	46.0
	$\text{SUM-GAN}_{\text{dpp}}$[30]	无	51.7
	Li 等[16]	有	52.7
	vsLSTM[17]	有	54.2
	SUM-LSTM(本文)	有	44.1
	SUM-attDecoder(本文)	有	52.9

表 3 数据增强下的性能对比

Tab.3 Performance comparison under data augmentation

数据集	算法	F 值	
		未使用数据增强	数据增强
SumMe	vsLSTM	37.6	41.6
	SUM-attDecoder	38.2	44.0
TVSum	vsLSTM	54.2	57.9
	SUM-attDecoder	52.9	58.9

与基准算法 SUM-LSTM 相比较,如表 2 所示,解码器引入视觉注意力机制后模型性能有了显著的提升. 具体地,在 SumMe 数据集上提升了 3.1%,而在 TVSum 上提升效果更显著,为 8.7%. 由此可见,本文提出的视觉注意力机制极大地改善了编解码器模型的性能,特别是在规模较大的数据集上,提升效果较为明显.

进一步,与文献[17]类似,笔者验证了所提 SUM-attDecoder 方法在数据增强情况下的性能,如表 3 所示. 具体地,随机选取数据集中 20% 的数据用于测试,将剩下 80% 的数据和另外 3 种数据集共同构成训练集和验证集. 可以看出,所提 SUM-attDecoder 方法在数据增强后性能有较为显著的提升,且在 2 个数据集上性能分别高于 vsLSTM 方法 2.4% 和 1.0%,这说明 SUM-attDecoder 更适合于大规模数据集. 这是因为在深度学习模型的训练过程中,当训练数据较少时容易过拟合,导致模型预测的准确性不高. 但当数据较大时,性能就会有较大提升,尽管这几个数据集在内容和风格上存在差异,但它们依然为模型的学习提供了更多的摘要标注数据,有利于深度模型的学习,因此最后生成的视频摘要的质量有所提高.

另外,笔者还分析了注意力范围对性能的影响. SUM-attDecoder 方法在预测每个视频帧重要性分数时,利用了历史的解码信息,通过融合历史解码序列作为注意力信息有效地指导当前视频帧重要性分数的预测,提升了模型预测的准确性. 而在模型的预测中不同时间跨度的相邻视频帧对当前时刻预测的指导作用不同,为此本文研究了注意力范围(视频样本序列长度)5 ~ 20 变化时对性能的影响,如图 4 所示.

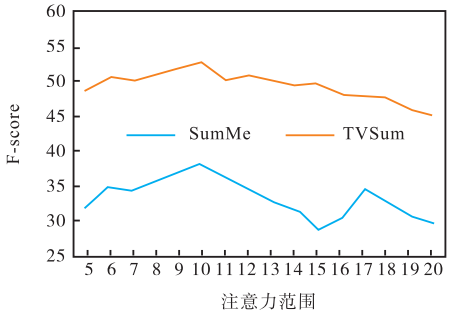

图 4 不同注意力范围对性能的影响

Fig.4 Influence of different attention scales on performance

可以看出,不同的注意力范围对性能影响不同,随着数值的增大性能会有所提升,在数值为 10 时达到顶峰,然后下降. 笔者分析最优的注意力范围为 10 的原因与镜头长度有关. 因为实验中 KTS 算法分割出的镜头平均长度为 10,通常同一个镜头的视频帧关联性较大,邻近视频帧的指导作用更大,而不同镜头视频帧的关联性较小. 从总体趋势可以看出注意力范围过大和过小时模型的性能会较差,注意力范围过小时关注的邻近视频帧较少,指导信息稍显不足;过大时关注的邻近视频帧较多,不同镜头的视频帧会干扰模型的预测,致使生成视频摘要的质量较低.

4 结 语

本文设计了一种新颖的基于解码器的视觉注意力机制,通过与现有的编解码器框架结合起来,提出一种新的有监督视频摘要算法. 所提算法将视频摘要看作是视频序列到重要性分数序列的预测问题,利用 LSTM 网络对该问题进行建模,并在解码器部分引入了视觉注意力机制,有效地利用历史的解码信息,提升了模型预测的准确性. 大量的实验结果证明了所提算法的有效性和先进性,并且分析了数据增强、视觉注意力模型、注意力范围对本文算法的影响.

有监督学习的方法依赖于大量的人工标注数据，现有的视频摘要数据集规模较小，导致训练模型不够充分. 今后的研究方向是如何在有限的标签数据下提升模型的泛化性，可以借鉴迁移学习[35]等技术.

参考文献：

[1] 王　娟，蒋兴浩，孙锬锋. 视频摘要技术综述[J]. 中国图象图形学报，2014，19(12)：1685-1695.
Wang Juan，Jiang Xinghao，Sun Tanfeng. Review of video abstraction[J]. *Journal of Image and Graphics*，2014，19(12)：1685-1695(in Chinese).

[2] de Avila S E F，Lopes A P B. VSUMM：A mechanism designed to produce static video summaries and a novel evaluation method[J]. *Pattern Recognition Letters*，2011，32(1)：56-68.

[3] Furini M，Geraci F，Montangero M，et al. STIMO：Still and moving video storyboard for the web scenario[J]. *Multimedia Tools and Applications*，2010，46(1)：47-69.

[4] Kuanar S K，Panda R，Chowdhury A S. Video key frame extraction through dynamic delaunay clustering with a structural constraint[J]. *Journal of Visual Communication and Image Representation*，2013，24(7)：1212-1227.

[5] Wu J，Zhong S H，Jiang J，et al. A novel clustering method for static video summarization[J]. *Multimedia Tools & Applications*，2017，76(7)：1-17.

[6] Ji Z，Zhang Y Y，Pang Y W，et al. Hypergraph dominant set based multi-video summarization[J]. *Signal Processing*，2018，148：114-123.

[7] Demir M，Bozma H I. Video summarization via segments summary graphs[C]//*IEEE International Conference on Computer Vision*. Santiago，Chile，2016：1071-1077.

[8] 冀　中，樊帅飞，基于超图排序算法的视频摘要[J]. 电子学报，2017，45(5)：1035-1043.
Ji Zhong，Fan Shuaifei. Video summarization with hypergraph ranking[J]. *Acta Electronica Sinica*，2017，45(5)：1035-1043(in Chinese).

[9] Panda R，Kuanar S K，Chowdhury A S. Scalable video summarization using skeleton graph and random walk[C]//*International Conference on Pattern Recognition.* Stockholm，Sweden，2014：3481-3486.

[10] Mei S，Guan G，Wang Z，et al. Video summarization via minimum sparse reconstruction[J]. *Pattern Recognition*，2015，48(2)：522-533.

[11] Panda R，Das A，Roy-Chowdhury A K. Video summarization in a multi-view camera network[C]// *International Conference on Pattern Recognition.* Cancun，Mexico，2016：2971-2976.

[12] Ji Z，Ma Y R，Pang Y W，et al. Query-aware sparse coding for multi-video summarization[EB/OL]. https://arxiv.org/abs/1707.04021，2017.

[13] Gong B，Chao W L，Grauman K，et al. Diverse sequential subset selection for supervised video summarization[C]//*Advances in Neural Information Processing Systems.* Montreal，Canada，2014：2069-2077.

[14] Zhang K，Chao W，Sha F，et al. Summary transfer：Exemplar-based subset selection for video summarization[C]//*IEEE Conference on Computer Vision and Pattern Recognition.* Las Vegas，USA，2016：1059-1067.

[15] Gygli M，Grabner H，van Gool L. Video summarization by learning submodular mixtures of objectives[C]//*IEEE Conference on Computer Vision and Pattern Recognition.* Boston，USA，2015：3090-3098.

[16] Li X，Zhao B，Lu X. A general framework for edited video and raw video summarization[J]. *IEEE Transaction on Image Processing*，2017，26(8)：3652-3664.

[17] Zhang K，Chao W L，Sha F，et al. Video summarization with long short-term memory[C]//*European Conference on Computer Vision.* Amsterdam，Netherlands，2016：766-782.

[18] Potapov D，Douze M，Harchaoui Z，et al. Category-specific video summarization[C]//*European Conference on Computer Vision.* Zurich，Sitzerland，2014：540-555.

[19] Yong J L，Ghosh J，Grauman K. Discovering important people and objects for egocentric video summarization[C]// *IEEE Conference on Computer Vision and Pattern Recognition.* Providence，USA，2012：1346-1353.

[20] Sutskever I，Vinyals O，Le Q V. Sequence to sequence learning with neural networks[C]//*Advances in Neural Information Processing Systems.* Montreal，Canada，2014：3104-3112.

[21] Ma Y F，Lu L，Zhang H J，et. al. A user attention model for video summarization[C]//*ACM Conference on Multimedia*. Juan les Pins，France，2002：533-542.

[22] 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.

[23] Bahdanau D，Cho K，Bengio Y. Neural machine trans-

lation by jointly learning to align and translate [C]//*International Conference on Learning Representations.* San Diego, USA, 2015: 1-15.

[24] Meng F, Lu Z, Wang M, et al. Encoding source language with convolutional neural network for machine translation[C]//*Annual Meeting of the Association for Computational Linguistics.* Beijing, China, 2015: 20-30.

[25] Chopra S, Auli M, Rush A M. Abstractive sentence summarization with attentive recurrent neural networks [C]//*Annual Meeting of the Association for Computational Linguistics.* Berlin, Germany, 2016: 93-98.

[26] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [C]//*International Conference on Machine Learning.* Lille, France, 2015: 2048-2057.

[27] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure[C]//*IEEE International Conference on Computer Vision.* Santiago, Chile, 2015: 4507-4515.

[28] Venugopalan S, Xu H, Donahue J, et al. Translating videos to natural language using deep recurrent neural networks[C]//*Annual Meeting of the Association for Computational Linguistics.* Baltimore, USA, 2014: 1494-1504.

[29] Li Y, Merialdo B. Multi-video summarization based on Video-MMR[C]//*International Workshop on Image Analysis for Multimedia Interactive Services.* Desenzano del Garda, Italy, 2010: 1-4.

[30] Mahasseni B, Lam M, Todorovic S. Unsupervised video summarization with adversarial LSTM networks [C]//*IEEE Conference on Computer Vision and Pattern Recognition.* Honolulu, USA, 2017: 1-10.

[31] Gygli M, Grabner H, Riemenschneider H, et al. Creating summaries from user videos[C]//*European Conference on Computer Vision.* Zurich, Switzerland, 2014: 505-520.

[32] Yang H, Wang B, Lin S, et al. Unsupervised extraction of video highlights via robust recurrent auto-encoders[C]// *IEEE International Conference on Computer Vision.* Santiago, Chile, 2015: 4633-4641.

[33] Song Y, Vallmitjana J, Stent A, et al. TVSum: Summarizing web videos using titles[C]//*IEEE Conference on Computer Vision and Pattern Recognition.* Boston, USA, 2015: 5179-5187.

[34] Zhao B, Xing E P, Quasi real-time summarization for consumer videos[C]//*IEEE Conference on Computer Vision and Pattern Recognition.* Columbus, USA, 2014: 2513-2520.

[35] Shao L, Zhu F, Li X. Transfer learning for visual categorization: A survey[J]. *IEEE Transactions on Neural Networks & Learning Systems*, 2015, 26(5): 1019-1034.

(责任编辑：王晓燕)