(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115190332 B
(45)授权公告日 2025.01.07

(21)申请号 202210801636.0

(22)申请日 2022.07.08

(65)同一申请的已公布的文献号
申请公布号 CN 115190332 A

(43)申请公布日 2022.10.14

(73)专利权人 西安交通大学医学院第二附属医院
地址 710004 陕西省西安市新城区西五路157号

(72)发明人 徐颂华 刘安然 周林韵 李宗芳 徐宗本

(74)专利代理机构 西安通大专利代理有限责任公司 61200
专利代理师 姚咏华

(51)Int.Cl.
H04N 21/234(2011.01)
H04N 21/44(2011.01)
H04N 21/488(2011.01)
H04N 5/278(2006.01)
G06V 10/762(2022.01)
G06V 10/764(2022.01)
G06V 10/82(2022.01)

(56)对比文件
CN 110929092 A,2020.03.27
CN 114627162 A,2022.06.14

审查员 熊艳

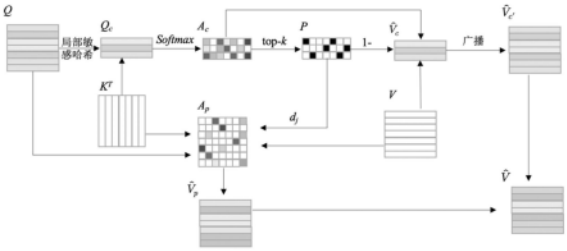
权利要求书3页 说明书9页 附图2页

(54)发明名称

一种基于全局视频特征的密集视频字幕生成方法

(57)摘要

本发明公开了一种基于全局视频特征的密集视频字幕生成方法,本发明通过自适应聚类的方法在只输入整段视频的情况下编码其全局特征,进而以端到端的方式指导事件定位和字幕生成,略去了先前模型利用先验阈值进行事件提案划分的步骤,从而在保证字幕生成准确性的条件下大大降低了计算复杂度。本发明在处理长序列特征时能够自适应地将相近的特征查询聚类进而降低冗余,节省内存。同时,作为传统Transformer中完整自注意力机制的快速近似,该方法在编码准确性方面也表现优异。



CN 115190332 B

1.一种基于全局视频特征的密集视频字幕生成方法,其特征在于,包括以下步骤:

运用预训练的动作识别网络提取视频的初级编码特征;

对初级编码特征进行处理,确定视觉中心和权重后再进行若干层堆叠,得到全局特征编码;

将全局特征编码作为指导,使用并行多头解码器来进行事件个数预测、事件定位以及字幕生成,最终生成视频字幕;事件个数预测采用事件个数预测头,具体方法如下:

将事件查询特征 $\{\tilde{e}_i\}_{i=1}^{L}$ 压缩为向量,然后运用全连接层预测一个固定长度的向量 $\rho \in \mathbb{R}^{L}$,其中每一个元素代表事件个数为事件查询特征的概率;

在推理阶段,选择置信度位于前$L_{\inf}$的提案作为最终的事件划分结果,每个事件生成字幕的置信度得分可以通过下式获得:

$$c_i = c_i^{loc} + \frac{\mu}{B_i^{\gamma}} \sum_{t=1}^{B_i} \log(c_{it}^{cap})$$

其中,c_{it}^{cap} 表示在第i个事件中生成第t个目标单词的概率,γ 为调制因子,μ为平衡因子,用来削弱字幕长度对置信度得分的影响;

事件定位采用事件提案定位头,具体方法如下:

事件提案定位头旨在对每个事件级特征生成框预测以及进行二分类,框预测的作用是为每个事件级特征预测其起始位置;二分类则为每个事件查询预测其前景置信度,这两部分预测都是将多层感知机运用在事件级特征 $\{\tilde{e}_i\}_{i=1}^{L}$ 上得到的:

$$(t_i^s, t_i^e) = \mathrm{MLP}_i^t(\tilde{e}_i)$$

$$c_i^{loc} = \mathrm{MLP}_i^c(\tilde{e}_i)$$

经过事件提案定位头,得到一组元组 $\{t_i^s, t_i^e, c_i^{loc}\}_{i=1}^{L}$ 来表示检测到的事件,其中 t_i^s, t_i^e 表示事件提案的起止时间,c_i^{loc} 表示对于事件特征 $\tilde{e}_i$ 的定位置信度;

字幕生成采用字幕生成头,具体方法如下:

对于初步编码的视频的初级特征F,为了提取其不同尺度的特征,将F输入ResNet并提取该网络C_3到C_5阶段的输出,记为 $\{x^m\}_{m=1}^{M-1}$,其中$M = 4$,第M个尺度的视频特征图是通过将一个卷积核为3×3,步长为2的卷积应用于C_5阶段的输出得到;

将语义查询$h_{i,t-1}$和事件级特征 $\tilde{e}_i$ 拼接作为查询,$h_{i,t-1}$表示字幕生成LSTM中的隐藏特征,对每个尺度的初级特征生成D个参考点,流程如下:

$$\tilde{x}_i = \mathrm{DSA}([h_{i,t-1}, \tilde{e}_i], g_i, X) = \bigcup_{m=1}^{M} \bigcup_{d=1}^{D} A_{imd} W x^m \tilde{g}_{imd}$$

$$\tilde{g}_{imd} = \phi_m(g_i) + \triangle g_{idm}$$

其中g_i直接由线性映射和sigmoid激活函数作用在查询 $[h_{i,t-1}, \tilde{e}_i]$ 上得到,它表示归一化的参考点的坐标,即$g_i \in [0,1]^2$,ϕ_m 将归一化参考点映射到对应尺度的特征图上,$\triangle g_{idm}$ 表示采样偏移量,A_{imd} 代表对于第i个语义事件查询,采样点d在m尺度上的注意力。

2.根据权利要求1所述的一种基于全局视频特征的密集视频字幕生成方法,其特征在于,提取视频的初级编码特征通过C3D 模型、双流网络结构或时间敏感视频编码器。

3.根据权利要求1所述的一种基于全局视频特征的密集视频字幕生成方法,其特征在于,得到全局特征编码的具体方法如下:

使用局部敏感哈希方法对初级编码特征中的视频特征进行处理,确定视频特征的视觉中心;

查询每一组视频特征,得到具有最高关注度的前k个视频特征键并确定权重;

重复上述步骤对所有编码的视频特征赋予权重,得到全局特征编码。

4.根据权利要求3所述的一种基于全局视频特征的密集视频字幕生成方法,其特征在于,使用局部敏感哈希方法对初级编码特征中的视频特征进行处理的具体方法如下:

计算每个初级编码特征中视频特征查询的哈希值;

将欧几里得局部敏感哈希作为哈希函数:

$$h(Q_j)=\lfloor\frac{\mathbf{a}\cdot Q_j+b}{r}\rfloor$$

其中,Q_j是Q的分量,r是超参数,a和b是随机变量,满足$\mathbf{a}=\left(a_1,\ a_2,\ \ldots,\ a_{D_q}\right)$,且$a_i\sim\mathcal{N}(0,1),b\sim\mathcal{U}(0,r)$,应用到H个LSH,得到每个视频分量的哈希值:

$$h(Q_j)=\sum_{i=0}^{H-1}B^i h_i(Q_j)$$

其中,B为常数;

设$Q_c\in\mathbb{R}^{U\times D_q}$是具有相同哈希值的视频特征查询的中心,$I_i$是类别索引,表示视频特征查询分量$Q_i$属于哪一组,第$j$组视觉中心$Q_{c(j)}$表示成下式:

$$Q_{c(j)}=\frac{\sum_{i,I_i=j}Q_i}{\sum_{i,I_i=j}1}$$

相应的集群注意力矩阵 $A_c\in\mathbb{R}^{U\times N}$ 按照如下方式得到:

$$A_c=\mathrm{softmax}(\frac{Q_cK^T}{\sqrt{D_q}})。$$

5.根据权利要求3所述的一种基于全局视频特征的密集视频字幕生成方法,其特征在于,得到具有最高关注度的前k个视频特征键并确定权重的具体方法如下:

设$P\in\{0,1\}^{U\times C}$是一组指示向量,其中$P_{ji}=1$当且仅当第i个视频特征键是第j组的关注度位于前k个的键之一,否则为0;

通过这种方式将在第j组中对关注度排在前k个的键和其它键分开并为它们计算如下的注意力系数:

$$d_j=\sum_{i=1}^{N}P_{ji}A_{c(ji)}$$

按照上述方式改进之后的注意力矩阵表示成:

$$A_{p(ir)}=\begin{cases}\frac{d_j\exp(Q_iK_r^T)}{\sum_{l=1}^{N}P_{jl}\exp(Q_iK_l^T)} & P_{jr}=1\\ A_{c(jr)} & \text{otherwise}\end{cases}$$

其中i表示的是第j个视频特征组中包含的第i个视频特征查询;

新的视频特征值 $\hat{V}_i$ 可以被分成如下两个部分:

$$\hat{V}_i = \hat{V}_{p(i)} + \hat{V}_{c'(i)}$$

其中,

$$\hat{V}_{p(i)} = \sum_{r=1}^{N} P_{jr} A_{p(ir)} V_r$$

$$\hat{V}_{c(j)} = \sum_{r=1}^{N} (1 - P_{jr}) A_{c(jr)} V_r$$

其中 $\hat{V}_{c'(i)}$ 由 $\hat{V}_{c(j)}$ 广播得到。

6.根据权利要求3所述的一种基于全局视频特征的密集视频字幕生成方法,其特征在于,得到全局特征编码的具体方法如下:

经过J层堆叠的包含自适应聚类注意的编码层,提取视频最终的全局特征编码 $S = \{s_1, \ldots, s_N\}$,所得到的全局视频特征不仅包含整段视频的背景信息,还应具有事物敏感性和事件敏感性。

一种基于全局视频特征的密集视频字幕生成方法

技术领域

[0001] 本发明属于视频特征提取领域,具体涉及一种基于全局视频特征的密集视频字幕生成方法。

背景技术

[0002] 随着多媒体平台的快速发展,越来越多的人们习惯从视频中获取信息。平均每天有数以千万计的视频被上传到互联网,而审核这些视频会消耗大量的时间。因此为视频自动生成描述性字幕的工作变得十分有价值,这不仅可以大大减少视频审核的时间,还可以借助语音朗读软件为视障患者获取信息。但是通常一个视频中包含多个相互关联的事件,只为视频生成单个的简短描述会丢失大量的信息,因此密集视频字幕生成任务应运而生。总的来说,该任务旨在对视频包含的每个事件进行定位并为其生成对应的字幕,整个过程主要包括两个子任务,即事件定位和字幕生成。而一个有竞争力的密集视频字幕生成模型应该在两个子任务上均具有良好的表现。

[0003] 现有的工作通常采用“事件定位-字幕生成”的串联式两阶段方案,该方案通常需要引入先验阈值对众多事件提案进行筛选,从而不可避免地增加了计算量和内存消耗;另外,该方案所生成的字幕质量严重依赖于事件定位的准确性,导致模型的性能很不稳定。

发明内容

[0004] 本发明的目的在于克服上述不足,提供一种基于全局视频特征的密集视频字幕生成方法,能够确保生成的视频字幕准确性的前提下尽可能提升计算效率。

[0005] 为了达到上述目的,本发明包括以下步骤:

[0006] 运用预训练的动作识别网络提取视频的初级编码特征;

[0007] 对初级编码特征进行处理,确定视觉中心和权重后再进行若干层堆叠,得到全局特征编码;

[0008] 将全局特征编码作为指导,使用并行多头解码器来进行事件个数预测、事件定位以及字幕生成,最终生成视频字幕。

[0009] 提取视频的初级编码特征通过C3D 模型、双流网络结构或时间敏感视频编码器。

[0010] 得到全局特征编码的具体方法如下:

[0011] 使用局部敏感哈希方法对初级编码特征中的视频特征进行处理,确定视频特征的视觉中心;

[0012] 查询每一组视频特征,得到具有最高关注度的前k个视频特征键并确定权重;

[0013] 重复上述步骤对所有编码的视频特征的赋予权重,得到全局特征编码。

[0014] 使用局部敏感哈希方法对初级编码特征中的视频特征进行处理的具体方法如下:

[0015] 计算每个初级编码特征中视频特征查询的哈希值;

[0016] 将欧几里得局部敏感哈希作为哈希函数:

[0017] $$h(Q_j)=\lfloor\frac{\mathbf{a}\cdot Q_j+b}{r}\rfloor$$

[0018] 其中，Q_J是Q的分量，r是超参数，a和b是随机变量，满足$\mathbf{a}=(a_1,\ a_2,\ \ldots,\ a_{D_q})$，且$a_i\sim\mathcal{N}(0,1),b\sim\mathcal{U}(0,r)$，应用到H个LSH，得到每个视频分量的哈希值：

[0019] $$h(Q_j)=\sum_{i=0}^{H-1}B^i h_i(Q_j)$$

[0020] 其中，B为常数；

[0021] 设$Q_c\in\mathbb{R}^{U\times D_q}$是具有相同哈希值的视频特征查询的中心，$I_j$是类别索引，表示视频特征查询分量$Q_i$属于哪一组，第$j$组视觉中心$Q_{c(j)}$表示成下式：

[0022] $$Q_{c(j)}=\frac{\sum_{i,I_i=j}Q_i}{\sum_{i,I_i=j}1}$$

[0023] 相应的集群注意力矩阵$A_c\in\mathbb{R}^{U\times N}$按照如下方式得到：

[0024] $$A_c=\mathrm{softmax}(\frac{Q_cK^T}{\sqrt{D_q}})$$

[0025] 得到具有最高关注度的前k个视频特征键并确定权重的具体方法如下：

[0026] 设$P\in\{0,1\}^{U\times C}$是一组指示向量，其中$P_{ji}=1$当且仅当第i个视频特征键是第j组的关注度位于前k个的键之一，否则为0；

[0027] 通过这种方式将在第j组中对关注度排在前k个的键和其它键分开并为它们计算如下的注意力系数：

[0028] $$d_j=\sum_{i=1}^{N}P_{ji}A_{c(ji)}$$

[0029] 按照上述方式改进之后的注意力矩阵表示成：

[0030] $$A_{p(ir)}=\begin{cases}\frac{d_j\exp(Q_iK_r^T)}{\sum_{l=1}^{N}P_{jl}\exp(Q_iK_l^T)} & P_{jr}=1\\ A_{c(jr)} & \text{otherwise}\end{cases}$$

[0031] 其中i表示的是第j个视频特征组中包含的第i个视频特征查询；

[0032] 新的视频特征值$\hat{V}_i$可以被分成如下两个部分：

[0033] $$\hat{V}_i=\hat{V}_{p(i)}+\hat{V}_{c'(i)}$$

[0034] 其中，

[0035] $$\hat{V}_{p(i)}=\sum_{r=1}^{N}P_{jr}A_{p(ir)}V_r$$

[0036] $$\hat{V}_{c(j)}=\sum_{r=1}^{N}(1-P_{jr})A_{c(jr)}V_r$$

[0037] 其中$\hat{V}_{c'(i)}$由$\hat{V}_{c(j)}$广播得到。

[0038] 得到全局特征编码的具体方法如下：

[0039] 经过J层堆叠的包含自适应聚类注意的编码层，提取视频最终的全局特征编码$S=\{s_1,\ldots,s_N\}$，所得到的全局视频特征不仅包含整段视频的背景信息，还应具有事物敏感性和事件敏感性。

[0040] 事件个数预测采用事件个数预测头，具体方法如下：

[0041] 将事件查询特征$\{\tilde{e}_i\}_{i=1}^{L}$压缩为向量，然后运用全连接层预测一个固定长度的向量$\rho\in\mathbb{R}^{L}$，其中每一个元素代表事件个数为该值的概率；

[0042] 在推理阶段，选择置信度位于前$L_{\inf}$的提案作为最终的事件划分结果，每个事件生成字幕的置信度得分可以通过下式获得：

[0043] $$c_i=c_i^{loc}+\frac{\mu}{B_i^{\gamma}}\sum_{t=1}^{B_i}\log(c_{it}^{cap})$$

[0044] 其中，c_{it}^{cap}表示在第i个事件中生成第t个目标单词的概率，γ为调制因子，μ为平衡因子，用来削弱字幕长度对置信度得分的影响。

[0045] 事件定位采用事件提案定位头，具体方法如下：

[0046] 事件提案定位头旨在对每个事件级特征生成框预测以及进行二分类，框预测的作用是为每个事件级特征预测其起始位置；二分类则为每个事件查询预测其前景置信度，这两部分预测都是将多层感知机运用在事件级特征$\{\tilde{e}_i\}_{i=1}^{L}$上得到的：

[0047] $$(t_i^s,t_i^e)=\mathrm{MLP}_i^t(\tilde{e}_i)$$

[0048] $$c_i^{loc}=\mathrm{MLP}_i^c(\tilde{e}_i)$$

[0049] 经过事件提案定位头，得到一组元组$\{t_i^s,t_i^e,c_i^{loc}\}_{i=1}^{L}$来表示检测到的事件，其中$t_i^s,t_i^e$表示事件提案的起止时间，$c_i^{loc}$表示对于事件特征$\tilde{e}_i$的定位置信度。

[0050] 字幕生成采用字幕生成头，具体方法如下：

[0051] 对于初步编码的视频的初级特征F，为了提取其不同尺度的特征，将F输入ResNet并提取该网络C_3到C_5阶段的输出，记为$\{x^m\}_{m=1}^{M-1}$，其中$M=4$，第M个尺度的视频特征图是通过将一个卷积核为3×3，步长为2的卷积应用于C_5阶段的输出得到；

[0052] 将语义查询$h_{i,t-1}$和事件级特征$\tilde{e}_i$拼接作为查询，$h_{i,t-1}$表示字幕生成LSTM中的隐藏特征，对每个尺度的初级特征生成D个参考点，，基本流程如下：

[0053] $$\tilde{x}_i=\mathrm{DSA}([h_{i,t-1},\tilde{e}_i],g_i,X)=\bigcup_{m=1}^{M}\bigcup_{d=1}^{D}A_{imd}Wx^m\tilde{g}_{imd}$$

[0054] $$\tilde{g}_{imd}=\phi_m(g_i)+\triangle g_{idm}$$

[0055] 其中g_i直接由线性映射和sigmoid激活函数作用在查询$[h_{i,t-1},\tilde{e}_i]$上得到，它表示归一化的参考点的坐标，即$g_i\in[0,1]^2$，ϕ_m将归一化参考点映射到对应尺度的特征图上，$\triangle g_{idm}$表示采样偏移量，A_{imd}代表对于第i个语义事件查询，采样点d在m尺度上的注意力。

[0056] 与现有技术相比，本发明通过自适应聚类的方法在只输入整段视频的情况下编码其全局特征，进而以端到端的方式指导事件定位和字幕生成，略去了先前模型利用先验阈值进行事件提案划分的步骤，从而在保证字幕生成准确性的条件下大大降低了计算复杂度。本发明在处理长序列特征时能够自适应地将相近的特征查询聚类进而降低冗余，节省

内存。同时,作为传统Transformer中完整自注意力机制的快速近似,该方法在编码准确性方面也表现优异。

附图说明

[0057] 图1为本发明中全局视频特征提取的流程图;

[0058] 图2为本发明中字幕生成头的整体流程图;

[0059] 图3为本发明的模型流程图。

具体实施方式

[0060] 下面结合附图对本发明做进一步说明。

[0061] 参见图1,基于自适应聚类的全局视频特征提取:

[0062] 本文通过自适应聚类的方法在只输入整段视频的情况下编码其全局特征,进而以端到端的方式指导事件定位和字幕生成,略去了先前模型利用先验阈值进行事件提案划分的步骤,从而在保证字幕生成准确性的条件下大大降低了计算复杂度。

[0063] 首先运用预训练的动作识别网络(C3D, TSN, TSP)来提取视频的初级编码特征$\{v_1,\ldots,v_N\}$。接着对这一初级编码特征进行处理得到有代表性的全局视频特征。

[0064] 运用插值将视频特征的时间维度重新缩放到N,从而得到视频的初级特征$F=\{f_i\}\in\mathbb{R}^{N\times H\times W}$。之后将初级特征展平并嵌入位置编码,作为包含自适应聚类编码器的

[0065] Transformer模型的输入:

[0066] $$F=\mathrm{CNN}(v_1,\ldots,v_N) \quad (1)$$

[0067] $$S=\mathrm{ACTAtt}(FW_Q,FW_K,FW_V) \quad (2)$$

[0068] 其中$W_Q\in\mathbb{R}^{HW\times D_q}$,$W_K\in\mathbb{R}^{HW\times D_q}$,$W_V\in\mathbb{R}^{HW\times D_v}$是可学习的参数,它们将视频的初级特征映射到编码器的输入空间。方便起见,不妨设所得到的视频特征查询为$Q\in\mathbb{R}^{N\times D_q}$,视频特征键$K\in\mathbb{R}^{N\times D_q}$,视频特征值$V\in\mathbb{R}^{N\times D_v}$。式(2)中的ACTAtt(•) 是本文的核心:基于自适应聚类注意力的编码器。主要思想是,首先将视频特征查询分到U个视频特征组,其中$U\ll N$。然后仅计算这些组的注意力,并对同一组的视频特征查询赋相同的注意力权重。进一步,为了使一些本应该获得较高关注度的视频特征键获得高度关注,还需要对关注度排在前k的键进行注意力重计算。总得来说,为了依据(2)式得到全局视频特征S,具体步骤如下(下面是对(2)式的具体解释,其中FW_Q代表Q,FW_K代表K,FW_V代表V):

[0069] 为了确定视频特征组,本文首先使用局部敏感哈希(LSH)方法对视频特征查询进行处理。考虑到LSH是解决最近邻搜索问题的强大工具:如果临近的向量能够以高概率获得相同的哈希值即落入相同的哈希桶中,而远距离的向量的哈希值不同,则称哈希方案局部敏感。因此通过控制哈希函数的相关参数和轮数,本文可以依据哈希值将所有距离小于ε的视频特征查询以大于p的概率分入同一视频特征组(哈希桶)中。具体来说,首先计算每个视频特征查询的哈希值,本文选择欧几里得局部敏感哈希作为哈希函数:

[0070] $$h(Q_j)=\lfloor\frac{\mathbf{a}\cdot Q_j+b}{r}\rfloor \quad (3)$$

[0071] 其中Q_j是Q的分量,r是超参数,a和b是随机变量,满足$a=(a_1,a_2,\cdots,a_{D_q})$且$a_i\sim\mathcal{N}(0,1)$,$b\sim\mathcal{U}(0,r)$, 应用$H$个LSH,最终得到的每个视频分量的哈希值如下:

[0072] $$h(Q_j)=\sum_{i=0}^{H-1}B^i h_i(Q_j) \quad (4)$$

[0073] 其中B是一个常数。从式 (3) 可以看出哈希函数实际上可以看成是一组具有随机法向量a和偏移量b的超平面,超参数r控制超平面的间距,r越大间距越大。而式 (3) 表明H个哈希函数将空间分成若干个单元格,落入同一单元格的向量将获得相同的哈希值。

[0074] 为了获得视觉中心,设$Q_c \in \mathbb{R}^{U\times D_q}$是具有相同哈希值的视频特征查询的中心,$I_i$是类别索引,表示视频特征查询分量$Q_i$属于哪一组。因此,第$j$组视觉中心$Q_{c(j)}$可以被表示成下式:

[0075] $$Q_{c(j)}=\frac{\sum_{i,I_i=j}Q_i}{\sum_{i,I_i=j}1} \quad (5)$$

[0076] 基于此,相应的集群注意力矩阵$A_c \in \mathbb{R}^{U\times N}$和视频特征值$\hat{V}_c \in \mathbb{R}^{U\times D_v}$可以按照如下方式得到:

[0077] $$A_c=\mathrm{softmax}(\frac{Q_c K^T}{\sqrt{D_q}}) \quad (6)$$

[0078] 进一步,对每一组视频特征查询找到具有最高关注度的前k个视频特征键并详细计算该部分的权重,剩余部分的权重依然按照上述聚类方式进行计算。

[0079] 具体来说,设$P \in \{0,1\}^{U\times C}$是一组指示向量,其中$P_{ji}=1$当且仅当第$i$个视频特征键是第$j$组的关注度位于前$k$个的键之一,否则为0。通过这种方式可以将在第$j$组中对关注度排在前$k$个的键和其它键分开并为它们计算如下的注意力系数(这样做的目的是保证前k个视频特征键和其余视频特征键所对应的值的注意力和为1):

[0080] $$d_j=\sum_{i=1}^{N}P_{ji}A_{c(ji)} \quad (8)$$

[0081] 式 (8) 实际上就是第j个视频特征组中关注度位于前k个视频特征键的总概率。那么按照上述方式改进之后的注意力矩阵可以表示成:

[0082] $$A_{p(ir)}=\begin{cases}\dfrac{d_j\exp(Q_i K_r^T)}{\sum_{l=1}^{N}P_{jl}\exp(Q_i K_l^T)} & P_{jr}=1\\ A_{c(jr)} & \text{otherwise}\end{cases} \quad (9)$$

[0083] 其中i表示的是第j个视频特征组中包含的第i个视频特征查询。换句话说,根据式 (6) 选择出每一个视频特征组关注度位于前k个的视频特征键,在注意力系数的缩放下,与该视频特征组中的每一个视频特征查询Q_i进行点积,再用$softmax$重新精细计算获得新的权值。对于不属于上述的视频特征键,依然按照式 (6) 仅在每一个视频特征组的视觉中心计算权重。总的来说,新的视频特征值$\hat{V}_i$可以被分成如下两个部分:

[0084] $$\hat{V}_i=\hat{V}_{p(i)}+\hat{V}_{c'(i)} \quad (10)$$

[0085] 其中,

[0086] $$\hat{V}_{p(i)}=\sum_{r=1}^{N}P_{jr}A_{p(ir)}V_r \quad (11)$$

[0087] $$\hat{V}_{c(j)}=\sum_{r=1}^{N}(1-P_{jr})A_{c(jr)}V_r \quad (12)$$

[0088] 其中$\hat{V}_{c'(i)}$由$\hat{V}_{c(j)}$广播得到。

[0089] 基于此,模型在每一个编码层中都对编码的视频特征执行上述操作,便可以得到一个具有代表性的视频全局特征。

[0090] 设经过J层堆叠的包含自适应聚类注意的上述编码层,模型就可以提取视频最终的全局特征编码$S=\{s_1,\ldots,s_N\}$。所得到的全局视频特征不仅包含整段视频的背景信息,还应具有事物敏感性和事件敏感性。

[0091] 参见图3,并行多头解码器将上面得到的全局视频特征S作为指导,使用并行多头解码器来同时进行事件个数预测、事件定位以及字幕生成三个下游子任务,从而促进子任务的交互并最终为视频生成准确的密集视频字幕描述。具体来说,本文的解码器并没有对输入的事件查询进行顺序递归处理,而是并行处理L个可学习的事件查询,旨在直接从以L个可学习嵌入为条件的带有丰富聚类信息的全局视频特征中查询事件级特征。若初始化的可学习事件查询表示为$E=\{e_i\}_{i=1}^{L}$,则在每一层解码层中注意力的计算流程可以表示为:

[0092] $$E'=\{e_i'\}_{i=1}^{L}=\mathrm{Att}(EW_Q',EW_K',EW_V') \quad (13)$$

[0093] $$\tilde{E}=\{\tilde{e}_i\}_{i=1}^{L}=\mathrm{Att}(E'\tilde{W}_Q,S\tilde{W}_K,S\tilde{W}_V) \quad (14)$$

[0094] 其中,Att(•)是自注意力机制。需要说明的是,式 (14) 中的键和值均来自编码器输出的全局视频特征S,解码层中自注意力的输出作为查询,本文称该注意力机制为交叉注意力机制。简单起见,这里依然只描述了一层解码层中的注意力部分,设经过J层解码层的迭代细化所得到的Transformer解码器的输出$\{\tilde{e}_i\}_{i=1}^{L}$即为运用全局视频特征指导的事件级特征。

[0095] 事件提案定位头

[0096] 事件提案定位头旨在对每个事件级特征生成框预测以及进行二分类。具体来说,框预测的作用是为每个事件级特征预测其起始位置;二分类则为每个事件查询预测其前景置信度,这两部分预测都是将多层感知机运用在事件级特征$\{\tilde{e}_i\}_{i=1}^{L}$上得到的:

[0097] $$(t_i^s,t_i^e)=\mathrm{MLP}_i^t(\tilde{e}_i) \quad (15)$$

[0098] $$c_i^{loc}=\mathrm{MLP}_i^c(\tilde{e}_i) \quad (16)$$

[0099] 这样,经过事件提案定位头,模型可以得到一组元组$\{t_i^s,t_i^e,c_i^{loc}\}_{i=1}^{L}$来表示检测到的事件,其中$t_i^s,t_i^e$表示事件提案的起止时间,$c_i^{loc}$表示对于事件特征$\tilde{e}_i$的定位置信度。

[0100] 参见图2,字幕生成头

[0101] 对于初步编码的视频的初级特征F,为了提取其不同尺度的特征,将F输入ResNet并提取该网络C_3到C_5阶段的输出,记为$\{x^m\}_{m=1}^{M-1}$,其中$M=4$,第M个尺度的视频特征图是通过将一个卷积核为3×3,步长为2的卷积应用于C_5阶段的输出得到的。尽管不同尺度的特征图的分辨率不同,但是可以通过1×1的卷积操作将它们的通道数转换成256。这样就得到了用来对字幕生成头进行视觉信息补充的多尺度视频初级特征,记为$X=\{x^m\}_{m=1}^{M}$。需要说明的是,这里没有选择经过自适应聚类编码器处理的全局视频特征作为视觉信息补充,原因是在没有经过聚类的特征上能够采样到更加丰富的视觉信息。

[0102] 进一步，当生成第i个事件查询的第t个单词时，首先需要对每个尺度的特征生成D个采样点，基本流程如下：

[0103] $$\tilde{x}_i = \mathrm{DSA}([h_{i,t-1}, \tilde{e}_i], g_i, X) = \bigcup_{m=1}^{M} \bigcup_{d=1}^{D} A_{imd} W x^m \tilde{g}_{imd} \quad (17)$$

[0104] 其中，

[0105] $$\tilde{g}_{imd} = \phi_m(g_i) + \triangle g_{idm} \quad (18)$$

[0106] 具体来说，将语义查询$h_{i,t-1}$和事件级特征$\bar{e}_i$拼接作为查询，这里$h_{i,t-1}$表示字幕生成LSTM中的隐藏特征。然后根据式 (17)，对每个尺度的初级特征生成D个参考点，其中g_i直接由线性映射和sigmoid激活函数作用在查询$[h_{i,t-1}, \tilde{e}_i]$上得到，它表示归一化的参考点的坐标，即$g_i \in [0,1]^2$。ϕ_m直接将归一化参考点映射到对应尺度的特征图上，$\triangle g_{idm}$表示采样偏移量，A_{imd}代表对于第i个语义事件查询，采样点d在m尺度上的注意力，这二者都是通过将线性投影作用到语义事件查询上得到的。

[0107] 这样，本文依据语义和事件查询，就可以得到在不同尺度的初级视频特征上采样的视频视觉信息的补充$\tilde{x}_i \in \mathbb{R}^{DM \times 256}$。接下来，根据软注意力的思想，可以对这些视觉信息采样点依照语义和事件查询进行加权处理：

[0108] $$a_{i,jt} = w_a^T \tanh(W_x \tilde{x}_j + W_{ha}[h_{i,t-1}, \tilde{e}_i]) \quad (19)$$

[0109] $$\alpha_{it} = \mathrm{softmax}(a_{it}) \quad (20)$$

[0110] 其中w_a，W_x，W_{ha}都是可学习的参数，$\tilde{x}_j \in \mathbb{R}^{256}$表示补充视觉信息的每一个分量。于是，加权后的视觉上下文特征可以表示为：

[0111] $$z_{it} = \sum_{j=1}^{DM} \alpha_{i,jt} \tilde{x}_j \quad (21)$$

[0112] 接下来，本文将补充的上下文视觉特征z_{it}，事件级特征$\bar{e}_i$以及之前的词嵌入$w_{i,t-1}$输入到LSTM中，得到时间步t的隐藏状态h_{it}，并进一步利用全连接层对下一个词w_{it}进行预测。那么对于第i个事件查询e_i，即可得到其对应的字幕$O_i = \{w_{i1}, \ldots w_{iB_i}\}$，其中$B_i$表示字幕的长度。

[0113] 事件个数预测头

[0114] 考虑到事件查询的个数L是一个人为设定的超参数，在实际的密集视频字幕生成任务中，并不需要对全部L个事件查询产生字幕。因为太多的事件会导致生成的字幕中有大量的重复，缺少可读性；而太少的事件又会导致重要信息的缺失。因此，本小节设计了一个事件个数预测头，旨在为每个视频预测一个合适的事件个数。

[0115] 具体来说，事件个数预测头包含一个最大值池化头和一个带有softmax激活的全连接层。首先，将事件查询特征$\{\tilde{e}_i\}_{i=1}^{L}$压缩为向量，然后运用全连接层预测一个固定长度的向量$\rho \in \mathbb{R}^L$，其中每一个元素代表事件个数为该值的概率。在推理阶段，选择置信度位于前L_{inf}的提案作为最终的事件划分结果，每个事件生成字幕的置信度得分可以通过下式获得：

[0116] $$c_i = c_i^{loc} + \frac{\mu}{B_i^{\gamma}} \sum_{t=1}^{B_i} \log(c_{it}^{cap}) \quad (22)$$

[0117] 其中，c_{it}^{cap}表示在第i个事件中生成第t个目标单词的概率，γ 为调制因子，μ为平衡

因子,用来削弱字幕长度对置信度得分的影响。

[0118] 为了证明本发明模型的优越性,本小节将本发明的模型与一些经典的密集视频字幕生成模型在事件定位准确性、字幕生成质量以及推理时间三方面进行对比。

[0119] 对事件定位准确性的评估:在早期的工作中,事件提案是通过预训练的模型提前生成的,不是端到端的结构。因此,在这一部分,本发明与经典的两阶段模型,也就是采用“定位-选择-描述”的管道式结构模型进行比较,以体现子任务并行策略的优点。具体来说:

[0120] 1.MT:作为本文的基线模型,该模型也基于Transformer的编-解码结构,首先将视频编码为适当的表示,事件提案解码器从带有不同锚点的编码中解码生成事件提案,字幕解码器根据提案解码器的输出生成字幕;

[0121] 2.MFT:将事件提案生成和字幕生成设计为一个循环网络,使得之前的字幕描述可以指导当前的事件提案划分;

[0122] 3.SDVC:考虑了事件的时间依赖性,并运用强化学习的手段在事件和情节连贯性两个方面进行两级奖励。

[0123] 那么,模型事件定位的精度、召回率以及F1结果如下表所示:

[0124] 表1 ActivityNet验证集上事件定位结果(使用C3D编码)

[0125]

方法	Re.					Pre.					F1
	0.3	0.5	0.7	0.9	平均	0.3	0.5	0.7	0.9	平均	
MFT	46.09	27.63	16.04	5.33	23.77	86.34	68.97	37.70	12.15	51.29	32.49
MT	52.95	34.73	25.78	9.85	30.83	90.23	74.01	42.15	11.22	54.40	39.36
SDVC	**93.32**	**75.19**	42.63	10.45	55.40	96.61	77.66	**44.85**	10.73	57.40	56.38
GDC_ori	89.04	72.66	45.30	**18.27**	**56.32**	96.89	78.00	41.02	**14.81**	57.68	57.00
GDC	90.25	72.43	**45.71**	15.79	56.05	**97.28**	**78.09**	43.31	14.10	**58.20**	**57.10**

[0126] 本发明和现有的方法采用的“事件定位-字幕生成”的串联式方案不同,本发明摒弃了这种方法,直接使用并行的方式输出事件提案定位,这可以大大减少方案中先验阈值的设置,而且比串联式方案更有效。从图中可以看出,本发明的事件提案定位结果远远超过了MFT和MT,并且能够和非端到端模型且拥有更多参数量的SDVC相当。特别地,当IOU阈值较高时,本发明模型的两个版本展示了更加有竞争力的结果和更加准确的定位性能。此外,当模型使用更精细的字幕生成头时,事件提案定位的平均精度也有所提升,这也表明两个并行头对应的子任务是相互影响相互促进的。

[0127] 为了分析全局视频特征在指导解码操作时的作用,本实验选取视频片段并通过全连接层标准化为帧级重要性输出,不同的事件以及事件中不同帧之间都有不同的权重,这为下游解码器生成事件提案和生成字幕提供了十分重要的指导。

[0128] 考虑到ActivityNet数据集上共包含203个动作类,并且字幕的性能可能跟动作的类别有关,本文还继续探索了在不同类别上GDC模型的表现。具体来说,本文选择了10个代表性类别,分别在上面评估得到GDC生成字幕的METEOR指标,并对比了在真实事件提案和预测事件提案上的性能,实验结果如表2所示:

[0129] 表2不同类别的动作上生成字幕的METEOR指标

[0130]

动作类型	预测提案	真实提案	动作类型	预测提案	真实提案
开碰碰车	8.87	16.19	跳高	7.90	8.11
拉小提琴	11.20	13.98	空手道	7.01	8.77
滑冰	9.01	11.12	潜水	8.02	10.71
打壁球	11.99	15.34	标枪	8.15	10.72
冲浪	8.35	13.20	跳尊巴	7.58	10.01

[0131] 从上表中可以看出,本发明在不同动作类型的视频数据上的表现是不同的。具体来说,对于一些大型的动作或者动作背景有一定的特异性的活动,比如开碰碰车、打壁球、滑冰等,本发明生成的字幕结果已经十分具有竞争力。但是对于一些更加小型以及细致的活动,比如空手道、跳尊巴等视频,本发明的表现略有下降。但是,基于表2的结果可以看出,通过更先进的视频特征提取模型对视频进行初步编码时捕捉更多的细粒度特征,GDC将有望实现更大的性能提升。需要说明的是,尽管对于运用自适应聚类提取全局视频特征的方法,修改部分超参数会带来字幕准确性的提升,但也会增大内存消耗,这将在下面的消融实验中进行更进一步的分析。

[0132] 综上所述,基于定量分析和定性分析的结果,GDC在两个数据集上均得到了十分有竞争力的表现,它比现有的密集视频字幕生成模型不论在事件定位还是字幕生成任务上都有十分显著的提升,这进一步证明了GDC的有效性。

[0133] 本发明在ActivityNet数据集和YouCookII数据集上将本发明与现有的密集视频字幕生成方法进行对比,实验结果表明不论在事件定位,字幕生成方面还是推理效率方面,本发明的模型都取得了领先。在预测提案以及真实提案上本发明都保证了其生成字幕的准确性。这些表现都进一步证实了本发明模型的有效性。另外,通过消融实验进一步分析了全局视频特征在下游任务中的行为。可视化的视频特征组进一步表明了全局视频特征的指导作用并增加了模型的可解释性。最后,分析了编码器的超参数H和r以及解码器中事件查询数量L对实验结果的影响,最终超参数的选择很好地平衡了模型在各项指标中的表现。

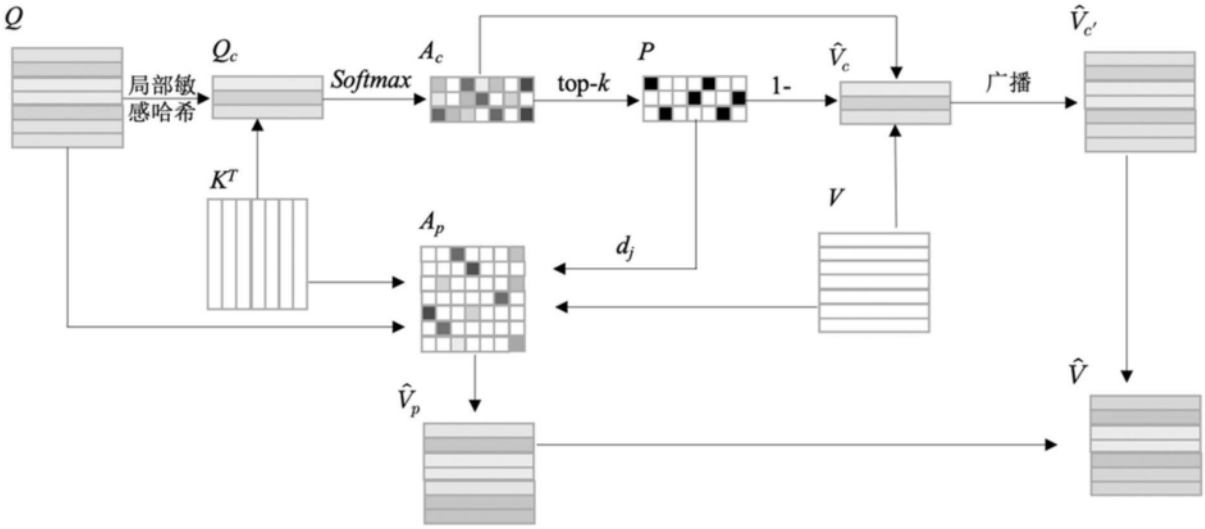

Q
局部敏感哈希
Q_c
Softmax
A_c
top-k
P
1-
$\hat{V}_c$
广播
$\hat{V}_{c'}$
K^T
A_p
d_j
V
$\hat{V}_p$
$\hat{V}$

图1

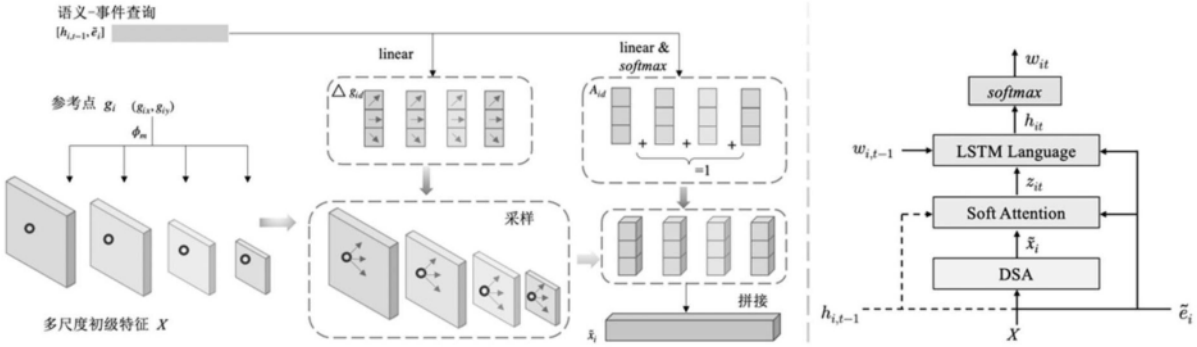

语义-事件查询
linear
linear & softmax
参考点 g_i
采样
拼接
多尺度初级特征 X
softmax
LSTM Language
Soft Attention
DSA

图2

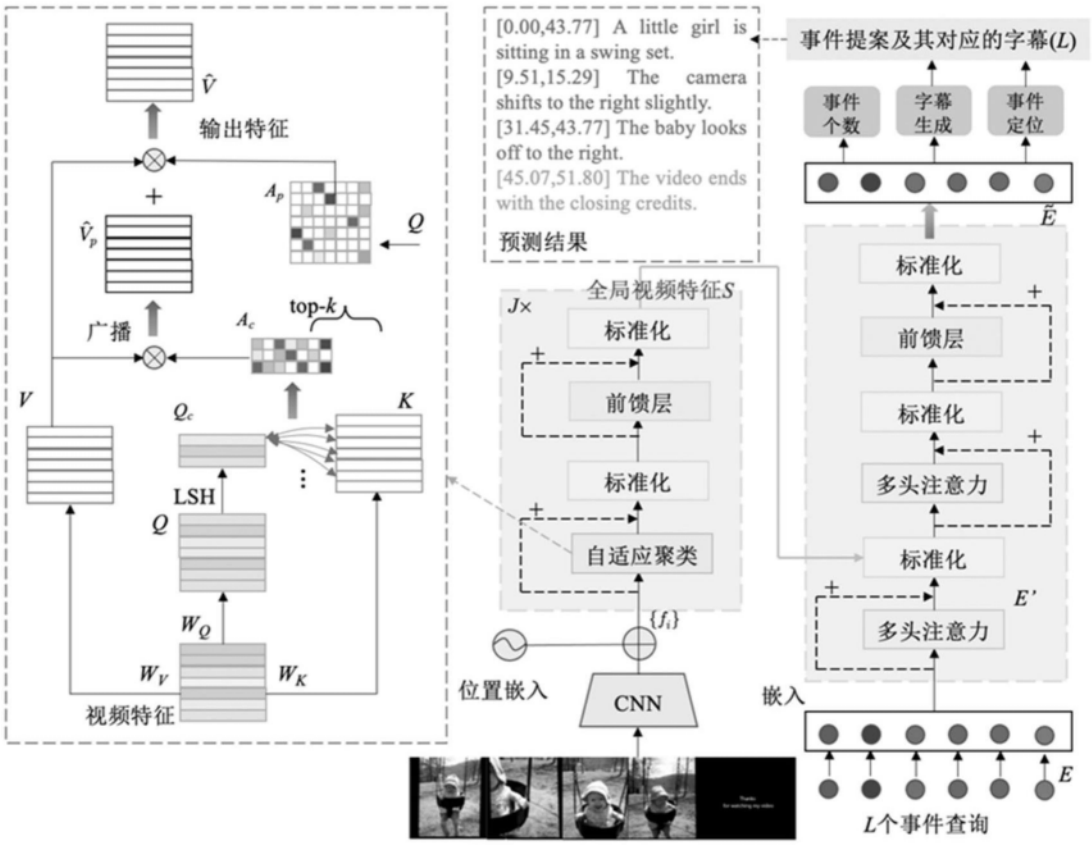

输出特征
广播
top-k
LSH
视频特征
[0.00,43.77] A little girl is sitting in a swing set.
[9.51,15.29] The camera shifts to the right slightly.
[31.45,43.77] The baby looks off to the right.
[45.07,51.80] The video ends with the closing credits.
预测结果
事件提案及其对应的字幕(L)
事件个数
字幕生成
事件定位
全局视频特征S
标准化
前馈层
多头注意力
自适应聚类
位置嵌入
CNN
嵌入
L个事件查询

图3