(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107979764 B
(45)授权公告日 2020.03.31

(21)申请号 201711273239.6

(22)申请日 2017.12.06

(65)同一申请的已公布的文献号
申请公布号 CN 107979764 A

(43)申请公布日 2018.05.01

(73)专利权人 中国石油大学(华东)
地址 266580 山东省东营市北二路271号

(72)发明人 吴春雷 魏燚伟 王雷全 褚晓亮 崔学荣

(74)专利代理机构 北京天奇智新知识产权代理有限公司 11340
代理人 陆军

(51)Int.Cl.
H04N 21/234(2011.01)
H04N 21/233(2011.01)
H04N 21/44(2011.01)
H04N 21/439(2011.01)
H04N 21/488(2011.01)
G06K 9/62(2006.01)
G06K 9/00(2006.01)

(56)对比文件
US 2015089518 A1,2015.03.26,
CN 105228033 A,2016.01.06,
CN 107038221 A,2017.08.11,
CN 107391646 A,2017.11.24,
CN 107391709 A,2017.11.24,

审查员 吴恂恂

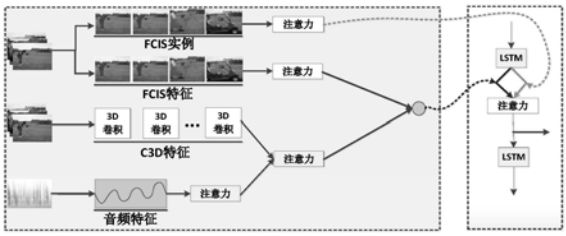
权利要求书3页 说明书8页 附图1页

(54)发明名称

基于语义分割和多层注意力框架的视频字幕生成方法

(57)摘要

本发明涉及基于语义分割与多模态注意力模型相结合的视频多字幕生成方法,包括:步骤1、从要生成字幕的视频中提取多帧图片;步骤2、利用全卷积实例感知语义分割模型,从视频提取某一反卷积层的特征信息;步骤3、提取视频的运动特征以及音频特征;步骤4、利用全卷积实例感知语义分割模型,从在步骤1中提取的图片中提取属性标签,其中,属性标签包含每帧图片中的物体信息;步骤5、并根据在前述步骤中提取的各个信息,生成不同模态的上下文矩阵,并对不同模态的上下文矩阵进行分层融合,生成融合后的上下文矩阵;步骤6、经由LSTM,通过多层感知机处理,得到作为字幕组成部分的单词;步骤7、将得到的所有单词进行串联组合,产生最终的字幕。



CN 107979764 B

1.一种基于全卷积语义分割与多模态注意力模型相结合的视频多字幕生成方法，包括以下步骤：

步骤1、从要生成字幕的视频中提取多帧图片；

步骤2、利用全卷积实例感知语义分割模型，从所述多帧图片提取某一反卷积层的特征信息；

步骤3、提取所述视频的运动特征以及音频特征；

步骤4、利用全卷积实例感知语义分割模型，从在所述步骤1中提取的图片中提取属性标签，其中，所述属性标签包含每一帧图片中的物体信息；

步骤5、并根据在前述步骤中提取的各个信息，生成不同模态的上下文矩阵，并对不同模态的上下文矩阵进行分层融合，生成融合后的上下文矩阵；

步骤6、初始化长短期记忆(LSTM)网络，将长短期记忆网络在前一时刻的隐藏层状态h^{t-1}和融合后的$context_t^{fuse}$传入长短期记忆网络，得到当前时刻的状态h^t，通过对h^t做多层感知机处理，得到作为字幕组成部分的单词$word_t$；

步骤7、判断是否在单词$word_t$中检测到停止标识，若检测到停止标识，则将得到的所有单词$word_t$进行串联组合，产生最终的字幕；若未检测到停止标识，则返回到步骤5。

2.根据权利要求1所述的方法，其中，在所述步骤3中，利用三维卷积神经网络提取所述视频的运动特征，利用小波变换来提取所述视频的音频特征。

3.根据权利要求1所述的方法，其中，所述不同模态的注意力模型包括属性模态注意力模型、视觉模态注意力模型、运动模态注意力模型、声音模态注意力模型。

4.根据权利要求3所述的方法，其中，在所述步骤5中，如下计算属性模态注意力模型的上下文矩阵$context_t^W$：

$$context_t^W = func^W(alpha_t^W, Words) \quad (2)$$

其中，Words＝Ins+$word_{t-1}$(1)

其中，在公式(1)中，Ins代表在所述步骤4中提取的属性标签，$word_{t-1}$代表模型上一时刻生成的单词，公式(1)的加号代表将属性标签和上一时刻生成的单词进行串联拼接的过程，Words代表所述拼接后生成的单词，

在公式(2)中，$context_t^W$的上标W代表对这些单词施加属性注意力操作，下标t代表时间序列，$func_w$表示属性注意力函数，$alpha_t^W$代表对于不同单词分配的权重。

5.根据权利要求4所述的方法，其中，在所述步骤5中，如下计算视觉模态注意力模型的上下文矩阵$context_t^V$：

$$context_t^V = func^V(alpha_t^V, Temporal\_features) \quad (8)$$

其中，Temporal_features＝Fcis_deconv(Multi_images) (7)

Multi_images＝Opencv(video,num) (6)

其中，Multi_images代表在步骤1中提取的多帧图片，Opencv是视频帧提取函数，video代表所要操作的视频，num代表需要提取帧的数量，Temporal_features代表在所述步骤2中提取的反卷积层的特征，Fcis_deconv代表反卷积层函数，$context_t^V$的上标V表示是对视觉信

息施加注意力操作。

6.根据权利要求5所述的方法，其中，在所述步骤5中，如下计算运动模态注意力模型的上下文矩阵 $context_t^M$：

$$context_t^M = func^M(alpha_t^M, Motion\_features) \quad (12)$$

其中，Motion_features＝C3D_conv(video) (11)

其中，Motion_features表示视频的运动特征矩阵，C3D_conv表示三维卷积神经网络函数，$func^M$表示运动注意力函数；其中：

$$alpha_t^M = \frac{\exp(\text{Motion\_features}_{i,t}, h^{t-1})}{\sum_{\kappa=1}^{L1}\exp(\text{Motion\_features}_{i,\kappa})} \quad (13)$$

其中，$alpha_t^M$ 表示运动注意力模型中运动矩阵第i个区域在t时刻的权重；$\text{Motion_features}_{i,k}$和$\text{Motion_features}_{i,t}$分别表示运动特征矩阵的不同位置，L1表示运动特征矩阵的列数。

7.根据权利要求6所述的方法，其中，在所述步骤5中，如下计算声音模态注意力模型的上下文矩阵 $context_t^A$：

$$context_t^A = func^A(alpha_t^A, Audio\_features) \quad (16)$$

其中，Audio_features＝MFCC(audio) (15)

其中，Audio_features表示声音特征矩阵，MFCC表示倒频谱系数函数，audio表示所述视频的音频数据，$func^A$表示声音注意力函数；

其中权重计算为：

$$alpha_t^A = \frac{\exp(\text{Audio\_features}_{i,t}, h^{t-1})}{\sum_{\kappa=1}^{L2}\exp(\text{Audio\_features}_{i,\kappa})} \quad (17)$$

其中，$alpha_t^A$ 表示声音注意力模型中声音矩阵第i个区域在t时刻的权重，$\text{Audio_features}_{i,k}$和$\text{Audio_features}_{i,t}$分别表示声音矩阵的不同位置，L2表示声音矩阵的列数。

8.根据权利要求7所述的方法，其中，在所述步骤5中，通过如下步骤，对不同模态的上下文矩阵进行分层融合，生成融合后的上下文矩阵：

步骤5-1、一层注意力融合，生成上下文矩阵 $\text{context}_t^{A\_M}$：

$$\text{context}_t^{A\_M} = \text{func}^{A\_M}\left(\text{beta}_t^{A\_M}, concentrate\left(\text{context}_t^A, \text{context}_t^M\right)\right) \quad (20)$$

其中，concentrate代表级联操作，$func^{M_A}$表示融合注意力函数，$\text{beta}_t^{A_M}$ 表示融合注意力模型中第i个区域在t时刻的权重；

$$\text{beta}_t^{A\_M} = \frac{\exp(A\_M\_context_{i,t}, h^{t-1})}{\sum_{k=1}^{L3}\exp(A\_M\_context_{i,k})} \quad (27)$$

其中，$A_M_context_{i,t}$表示所述第i个区域的 $context_t^A$ 和 $context_t^M$ 的级联，L3表示上下文矩阵的列数；

步骤5-2、二层注意力融合，生成融合后的上下文矩阵 $context_t^{fuse}$：

$$context_t^{fuse} = context_t^{A\_M} + context_t^V \quad (21)$$

其中公式(21)表示通过相加操作对不同的注意力模型进行融合。

9.根据权利要求8所述的方法，其中，在所述步骤6，如下得到作为字幕组成部分的单词 $word_t$：

$$h_{left}^t = LSTM\left(h^{t-1}, context_t^{fuse}\right) \quad (22)$$

$$h_{right}^t = LSTM\left(h^{t-1}, context_t^W\right) \quad (23)$$

$$beta^t = soft\max(nonlinearization(h_{left}^t), nonlinearization(h_{right}^t)) \quad (24)$$

$$h^t = beta^t * concat(h_{left}^t, h_{right}^t) \quad (25)$$

$$word_t = MLP(h^t) \quad (26)$$

其中，h_{left}^t 表示长短期记忆网络的左分支，h_{right}^t 表示LSTM的右分支，softmax表示回归函数，nonlinearization表示非线性化操作，MLP表示多层感知机函数。

基于语义分割和多层注意力框架的视频字幕生成方法

技术领域

[0001] 本发明涉及计算机视觉和自然语言处理的技术领域，特别是涉及基于计算机视觉的三维特征提取技术和语义分割技术、以及基于自然语言处理的时序模型技术，更具体地，涉及基于全卷积语义分割和多层注意力框架的视频字幕生成方法。

背景技术

[0002] 视频字幕生成指的是对一段视频自动生成自然语言描述。此类研究在人工智能和计算机视觉领域受到越来越多的关注。在当今社会，它具有非常广泛的应用，例如帮助盲人的日常生活，提高视频在线检索的质量等。除了相关应用之外，视频字幕生成技术对计算机视觉领域和跨模态技术的发展起到了巨大的推动作用。不同于单一的图像处理技术，视频字幕生成不仅要考虑到时间空间的相互协调，还要顾及到视频信息和语义信息的结合。

[0003] 现有的对视频字幕生成方法的研究主要分为两大方向，分别是基于更多模态融合的方法、以及优化传统注意力模型的方法。

[0004] 基于更多模态融合的方法以信息论为基础，尽可能地利用视频中的不同种类的信息，例如视频时空信息、分类信息和音频信息等。通过相关融合技术来提高生成字幕(描述)的质量。

[0005] 优化传统注意力模型的方法主要受到图片描述中软注意力模型的启发。考虑到视频描述的动态性、多样性等特点，通过改变注意力的施加方式及位置来提高生成字幕的质量。

[0006] 此外，一些科研院所提出了多模态融合技术，其不仅利用了不同的模态信息，还能有效的把不同信息融合在一起。

[0007] 相比于上述传统方法，多模态融合技术在准确度和个性化方面具有优点。同时，多模态融合技术还存在很多的不足。例如，由于视频的特性，导致视频需要提取大量的图片特征，而对每一张图片都进行大小调整会丢失图片的结构信息，并且提取三维卷积和二维卷积看似提取出了不同的特征，但由于卷积的权值共享因素导致了大量信息的重复提取。目前，虽然利用注意力机制提升了模态之间的融合效果，但对不同模态利用同一注意力操作并没有考虑模态之间存在差异，这会导致模态间的信息交叉干扰。

发明内容

[0008] 本发明的目的是针对现有技术的不足，提供一种基于完全卷积语义分割和多层注意力模型相结合的视频字幕生成方法。本发明首次实现了把语义分割技术利用到视频字幕生成当中。具体地，以语义分割代替传统的视觉特征，并通过融合语义分割产生的相关词汇来优化视频字幕的质量。

[0009] 根据本发明的实施例，提供了一种基于全卷积语义分割与多模态注意力模型相结合的视频多字幕生成方法，包括以下步骤：

[0010] 步骤1、从要生成字幕的视频中提取多帧图片；

[0011] 步骤2、利用全卷积实例感知语义分割模型，从所述视频提取某一反卷积层的特征信息；

[0012] 步骤3、提取所述视频的运动特征以及音频特征；

[0013] 步骤4、利用全卷积实例感知语义分割模型，从在所述步骤1中提取的图片中提取属性标签，其中，所述属性标签包含每一帧图片中的物体信息；

[0014] 步骤5、并根据在前述步骤中提取的各个信息，生成不同模态的上下文矩阵，并对不同模态的上下文矩阵进行分层融合，生成融合后的上下文矩阵；

[0015] 步骤6、初始化LSTM网络，将LSTM网络在前一时刻的隐藏层状态h^{t-1}和融合后的$context_t^{fuse}$传入LSTM网络，得到当前时刻的状态h^t，通过对h^t做多层感知机处理，得到作为字幕组成部分的单词$word_t$；

[0016] 步骤7、判断是否在单词$word_t$中检测到停止标识，若检测到停止标识，则将得到的所有单词$word_t$进行串联组合，产生最终的字幕；若未检测到停止标识，则返回到步骤5。

[0017] 由此，本发明提出了一种新的方法来生成视频描述，在各种普及的标准基准上表现出了较好的效果。与现有技术不同，本发明第一次提出了利用fcis(fully convolutional instance-aware semantic segmentation)与多层注意力相结合的方法，尽可能利用视频的有用信息，摒弃无用信息，并模拟现实情况提出动作与声音结合的方式。因此，本发明的方法不仅利用了fcis属性和特征突出化的有点，而且还科学地对不同模态施加注意力，让生成的句子(视频描述)更能够真实的反应视频的内容。本发明的方法能够极大地提高不同模态信息的利用率。

附图说明

[0018] 图1为根据本发明的实施例的基于全卷积语义分割和多层注意力框架的视频字幕生成方法的架构示意图；

[0019] 图2为本发明的实施例所采用的LSTM网络的结构示意图。

具体实施方式

[0020] 下面，结合附图对技术方案的实施作进一步的详细描述。

[0021] 本领域的技术人员能够理解，尽管以下的说明涉及到有关本发明的实施例的很多技术细节，但这仅为用来说明本发明的原理的示例、而不意味着任何限制。本发明能够适用于不同于以下例举的技术细节之外的场合，只要它们不背离本发明的原理和精神即可。

[0022] 另外，为了避免使本说明书的描述限于冗繁，在本说明书中的描述中，可能对可在现有技术资料中获得的部分技术细节进行了省略、简化、变通等处理，这对于本领域的技术人员来说是可以理解的，并且这不会影响本说明书的公开充分性。

[0023] 下面结合附图对具体实施方案进行详细描述。

[0024] 本发明的目的在于针对每一时刻生成的词，减小对上一次时刻词的依赖，已达到更准确的描述图像的效果。

[0025] 本发明分别采用C3D(三维卷积神经网络)、MFCC(倒谱系数)、fcis(全卷积语义分割)的全卷积来提取视频动作、时间和声音特征。并且本发明利用fcis的语义分割技术提取视频不同帧的属性。在t时刻，对于第一层注意力模型，主要对音频特征做注意力处理，对于

第二层注意力模型，分别针对不同性质的模态信息做注意力处理，对于第三层注意力模型，通过对生成的LSTM的状态做注意力处理。整个模型的架构如图1所示。

[0026] 图1为根据本发明的实施例的基于全卷积语义分割和多层注意力框架的视频字幕生成方法的架构示意图。如图1所示，FCIS特征(feature)代表对从视频中抽取的图像(帧)提取特征(对应上方的图像特征提取工作)，FCIS实例(Instance)代表从视频图像中提取的属性标签，C3D特征(feature)代表提取的三维特征。音频特征(Audio feature)代表提取的声音特征。LSTM代表长短期记忆网络。注意力(Attention)代表不同模态的注意力操作。从图1中可以看出，本发明利用了层次型的注意力方法，用不同层的注意力来编码不同的模态。从图1中还可看出，本发明结合了FCIS的卷积操作和标签提取操作。这正明确了本发明所提出的基于全卷积语义分割(FCIS)和多层注意力相结合的方法。

[0027] 根据本发明的实施例，提供了一种基于语义分割与多模态注意力模型相结合的视频多字幕生成方法，包括以下步骤(1)至(7)，下面逐一说明。

[0028] 步骤(1)：利用OPENCV库，从要生成字幕的视频中提取关键帧，并保存成图片格式，如.jpg格式；

[0029] 步骤(2)：利用全卷积实例感知语义分割模型(Fully Convolutional Instance-aware Semantic Segmentation)代替传统的Resnet(残差网络)模型，从视频提取某一卷积层的特征信息。

[0030] 步骤(3)：利用C3D(三维卷积神经网络)提取视频的空间(三维)特征。利用小波变换技术提取视频中包含的音频特征；

[0031] 其中，上述步骤(1)至(3)是独立执行的步骤；

[0032] 步骤(4)：利用全卷积实例感知语义分割模型，从在步骤(1)中保存的关键帧图片中提取提取属性标签。属性标签主要包含每一帧图像中的物体信息。如图片中有“人”这个物体，就会把“人”这个词存入属性标签中；

[0033] 简而言之，上述步骤(1)-(3)是构建编码结构，步骤(4)构建解码结构。编码是用预先规定的方法将文字、数字或其它对象编成数码，或将信息、数据转换成规定的电脉冲信号。编码是信息从一种形式或格式转换为另一种形式的过程。解码，是编码的逆过程。

[0034] 步骤(5)：创建时序引导LSTM网络。其作用主要分为两点，一是提供了对不同模态实施注意力的依据，二是作为字幕生成方法的主体引导框架。以LSTM网络的t-1时刻的隐藏层状态h_{t-1}^{1}产生的注意力模态的注意力向量，并将其与空间嵌入后的不同模态的特征相互结和，产生不同模态的上下文矩阵c_t，并根据视频的特性分层处理注意力模型，最后对不同层的上下文进行融合。其中，所述不同模态的注意力模型包括属性模态注意力模型、视觉模态注意力模型、动作模态注意力模型、声音模态注意力模型。

[0035] 步骤(6)：将h_{t-1}和融合后的$context_t^{fuse}$传入LSTM得到$\mathbf{h}_t^a$。通过对$\mathbf{h}_t^a$做多层感知机处理得到单词W_t。

[0036] 步骤(7)：判断是否在单词W_t中检测到停止标识，若是，则将得到的所有单词W_t进行串联组合，产生视频字幕；若不是，重复执行步骤(5)至(6)，直至检测到停止标识。

[0037] 在步骤(1)至(4)中，可使用现有方法提取不同的特征，为了使本说明书的描述不限于冗繁，在此不再详述。

[0038] 下面所以对步骤(5)至(6)的实现(公式)进行详解。

[0039] 所述步骤(5)的实现：

[0040] 一、属性模态注意力模型相关公式：

[0041] Words＝Ins+$word_{t-1}$ (1)

[0042] $$context_t^W = func^W(alpha_t^W, Words) \quad (2)$$

[0043] 公式(1)中，Ins代表在步骤(4)中提取的属性标签，$word_{t-1}$代表模型上一时刻生成的单词。而公式(1)的加号代表将属性标签和上一时刻生成的单词进行拼接的过程，Words代表所述拼接后生成的词(总和)。

[0044] 公式(2)中，$context_t^W$代表施加注意力后的上下文矩阵，其中的上标W代表“Word”，主要用来说明是对这些“Word”(属性标签和上一时刻生成的词)施加注意力操作，下标t代表时间。$func_w$表示属性注意力函数。alpha代表对于不同Words分配的权重，其是一个向量，若有n个Words，那它就有n维。

[0045] 下面是对属性注意力函数的说明。

[0046] 使用nlp(自然语言处理)中的embedding(空间嵌入)方法并结合非线性化过程，将属性标签Words转化成一个N*L维的向量表示：

[0047] $$\text{Words} = \{\{w_1^1 \ldots\ldots\ldots\ldots w_1^L\} \{w_2^1 \ldots\ldots\ldots\ldots w_2^L\} \ldots\ldots \{w_N^1 \ldots\ldots\ldots\ldots w_N^L\}\}$$

[0048] 其中，N为维单词的个数，L为空间嵌入后的属性标签的维度，空间嵌入是一个对向量从低维空间到高维空间转化的过程，如向量本身的维度为m经过空间嵌入后可以变为L，L的具体大小根据情况而定，对于属性矩阵的每个区域，属性注意力模型回归函数softmax根据属性矩阵Words和LSTM在t-1时刻的状态h_{t-1}产生权重向量$alpha_t^W$；

[0049] $$alpha_t^W = soft\max(Words, h_{t-1}) \quad (3)$$

[0050] 上标W代表word的意思，表示在对语义信息(属性标签)做注意力操作，而非其他的模态。

[0051] 并进行归一化处理：

[0052] $$alpha_t^W = \frac{\exp(word_{i,t}, h_{t-1})}{\sum_{\kappa=1}^{L}\exp(word_{i,\kappa})} \quad (4)$$

[0053] 其中，$alpha_t^W$表示属性注意力模型中属性标签矩阵第i个区域(其中i代表第i个单词对应的向量，而对应到属性标签矩阵中就是第i个区域)在t时刻的权重；$word_{i,k}$和$word_{i,t}$分别表示属性矩阵Words的不同位置。

[0054] 作为示例，经过属性注意力模型处理以后的属性上下文为$context_t^W$：

[0055] $context_t^W = alpha_t^W \times Words_t^W$ (5)

[0056] 二、视觉模态注意力模型相关公式：

[0057] Multi_images＝Opencv(video,num) (6)

[0058] Temporal_features＝Fcis_deconv(Multi_images) (7)

[0059] $context_t^V = func^V(alpha_t^V, Temporal\_features)$ (8)

[0060] 其中，Multi_images代表在步骤(1)中提取的多帧图片，比如一个视频有150帧，从中提取100帧，那么Multi_images就代表100张图片的总和。而Opencv是一种通用的视频帧提取工具。video代表所要操作的视频，num代表需要提取帧的数量。Temporal_features代表通过全卷积语义分割网络提取的反卷积层的特征。Fcis_deconv代表Fcis(Fully Convolutional Instance-aware Semantic Segmentation)反卷积层函数。$context_t^V$ 表示施加注意力的图像上下文。其中V代表visual，表示是对视觉信息施加注意力操作。

[0061] 其中，公式(6)表示利用现有opencv技术每几帧提取视频图片(关键帧)的过程。公式(7)表示利用Fcis的反卷积层来从多帧图片中提取反卷积特征，而t代表时刻序列，$func^V$ 代表视觉注意力函数，$alpha_t^V$ 代表对于不同帧分配的权重。

[0062] 下面说明视觉注意力函数。

[0063] 使用多张图片形成的时间特征作为特征输入，其被表示为一个N*L*D的三维矩阵(时间特征矩阵)，具体形式如下：

[0064] Temporal_features＝{T_1,T_2.....T_D}

[0065] 其中，T_i表示每一张图片的特征，其维度为(N,L)，D表示图片(关键帧)个数。

[0066] 对于时间特征矩阵Temporal_features的每一张图片的特征T_i，视觉注意力模型回归函数softmax根据时间特征矩阵Temporal_features和LSTM在t-1时刻的状态h_{t-1}产生权重向量$alpha_t^V$；

[0067] $$alpha_t^V = \frac{\exp(Temporal\_features_{i,t}, h_{t-1})}{\sum_{\kappa=1}^{L}\exp(Temporal\_features_{i,\kappa})} \quad (9)$$

[0068] 其中，$alpha_t^V$ 表示视觉注意力模型中图像矩阵第i个区域(其中i代表第i帧对应的向量，而对应到图像矩阵中就是第i个区域。在t时刻的权重；$Temporal_features_{i,k}$和$Temporal_features_{i,t}$分别表示图像矩阵的不同位置。

[0069] 经过视觉注意力模型处理以后的视觉上下文为$context_t^v$：

[0070] $context_t^v = alpha_t^v \times Temporal\_features_t^v$ (10)

[0071] 三、动作模态注意力模型相关公式：

[0072] Motion_features＝C3D_conv(video) (11)

[0073] $context_t^M = func^M(alpha_t^M, Motion\_features)$ (12)

[0074] 其中，公式(11)表示利用C3D(三维卷积神经网络)从目标视频提取三维特征，其中三维特征代表三维卷积特征，其是从一般的二维卷积特征发展而来，主要用来对视频中的动作提取特征(可参见Learning Spatiotemporal Features with 3D Convolutional

Networks)，公式(12)中的$func^{M}$表示动作注意力函数，$alpha_t^M$表示动作注意力模型中第i个区域在t时刻的权重，t代表时间序列；其中：

[0075] $$alpha_t^M = \frac{\exp(\text{Motion\_features}_{i,t}, h^{t-1})}{\sum_{\kappa=1}^{L}\exp(\text{Motion\_features}_{i,\kappa})} \quad (13)$$

[0076] 其中，$alpha_t^M$表示动作注意力模型中动作矩阵第i个区域在t时刻的权重；$Motion_features_{i,k}$和$Motion_features_{i,t}$分别表示动作矩阵的不同位置。

[0077] 经过动作注意力模型处理以后的动作上下文为$context_t^{M}$：

[0078] $$context_t^{M} = alpha_t^{M} \times \text{Motion\_features}_t^{M} \quad (14)$$

[0079] M代表motion的意思，表示在对动作信息做注意力操作，而非其他的模态。

[0080] 四、声音模态注意力模型相关公式：

[0081] Audio_features＝MFCC(audio) (15)

[0082] $$context_t^A = func^A(alpha_t^A, Audio\_features) \quad (16)$$

[0083] 其中，公式(15)表示利用MFCC(倒频谱系数)对目标声音提取音频特征，公式(16)的$func^{A}$表示声音注意力函数，$alpha_t^A$表示声音注意力模型中第i个区域在t时刻的权重，t代表时间序列；

[0084] 其中权重计算为：

[0085] $$alpha_t^A = \frac{\exp(\text{Audio\_features}_{i,t}, h^{t-1})}{\sum_{\kappa=1}^{L}\exp(\text{Audio\_features}_{i,\kappa})} \quad (17)$$

[0086] 其中，$alpha_t^A$表示声音注意力模型中声音矩阵第i个区域(第i个区域代表声音矩阵的第i行，没有实际意义)在t时刻的权重；$Audio_features_{i,k}$和$Audio_features_{i,t}$分别表示声音矩阵的不同位置。

[0087] 经过声音注意力模型处理以后的声音上下文为$context_t^{A}$：

[0088] $$context_t^{A} = alpha_t^{A} \times Audio\_features_t^{A} \quad (18)$$

[0089] A代表audio的意思，表示在对声音信息做注意力操作，而非其他的模态。

[0090] 五、注意力模型融合

[0091] 一层注意力融合公式：

[0092] $$A\_M\_context = concentrate(context_t^A, context_t^M) \quad (19)$$

[0093] $$context_t^{A\_M} = func^{A\_M}(beta_t^{A\_M}, A\_M\_context) \quad (20)$$

[0094] 其中公式(19)中的concentrate代表级联操作。公式(20)的$func^{M_A}$表示融合注意力函数，$beta_t^{A_M}$表示融合注意力模型中第i个区域在t时刻的权重，t代表时间序列；

[0095] $$\mathrm{beta}_t^{A\_M}=\frac{\exp(A\_M\_context_{i,t},h^{t-1})}{\sum_{k=1}^{L}\exp(A\_M\_context_{i,k})} \quad (27)$$

[0096] 二层注意力融合公式：

[0097] $$\mathrm{context}_t^{fuse}=\mathrm{context}_t^{A\_M}+context_t^{V} \quad (21)$$

[0098] 其中公式(21)表示通过相加操作对不同的注意力模型进行融合。

[0099] 所述步骤(6)的公式为：

[0100] $$\mathrm{h}_{\mathrm{left}}^{t}=LSTM\left(h^{t-1},context_t^{fuse}\right) \quad (22)$$

[0101] $$\mathrm{h}_{\mathrm{right}}^{t}=LSTM(h^{t-1},context_t^{W}) \quad (23)$$

[0102] $$\mathrm{beta}^{t}=soft\max(\mathrm{nonlinearization}(\mathrm{h}_{\mathrm{left}}^{t}),\mathrm{nonlinearization}(\mathrm{h}_{\mathrm{right}}^{t})) \quad (24)$$

[0103] $$h^{t}=beta^{t}*concat(h_{left}^{t},h_{right}^{t}) \quad (25)$$

[0104] $\mathrm{word}_t=\mathrm{MLP}(\mathrm{h}^t)$ (26)

[0105] 其中，h_{left}^{t}表示LSTM的左分支，$\mathrm{h}_{\mathrm{right}}^{t}$表示LSTM的右分支，LSTM表示长短时记忆网络，h_{t-1}表示LSTM的上一个状态，softmax表示回归函数，nonlinearization表示非线性化操作，h^t表示LSTM当前状态，MLP表示多层感知机，word_t表示求得的单词。

[0106] 本发明采用维度为K的one-hot向量来表示：

[0107] 模型在t时刻产生的单词word_t的向量的维度为1x K。其中K表示词典的大小。

[0108] 视频生成的句子用维度为C*K的向量W来表示：

[0109] $\mathrm{W}=\{\mathrm{w}_1,\ldots,\mathrm{w}_C\}$，$\mathrm{w}_i\in\mathrm{R}^K$

[0110] 其中K表示词典的大小，C表示产生的句子的长度(单词数量)(单词的数量?)。

[0111] 下面说明本发明的实施例所使用的LSTM网络。

[0112] 图2为本发明的实施例所采用的LSTM网络的结构示意图。LSTM是循环神经网的特殊形式，它成功解决了循环神经网络的梯度消失和梯度爆炸问题，LSTM的核心是它在每个步骤中的存储单元Cell，每个存储单元由三个Gate(输入门(Input Gate)、遗忘门(Forget Gate)、输出门(Output Gate))和一个cell单元组成。Gate可使用sigmoid激活函数，而input和cell state可使用tanh来转换。

[0113] 有关构造LSTM网络的具体方式、以及LSTM的Gates、Cell、输入变换和状态更新的具体定义，可从现有资料获得，这对于本领域的技术人员来说是熟知的。为了使本说明书的描述不限于冗繁，在此不再详述。

[0114] 数据集及实验结果：

[0115] 下面，选择流行的Youtube2Text和MSR-VTT dataset评估本发明的模型的性能Youtube2Text包含10000个视频片段(video clip)，被分为训练，验证和测试集三部分。每个视频片段都被标注了大概20条英文句子。此外，MSR-VTT还提供了每个视频的类别信息(共计20类)，这个类别信息算是先验的，在测试集中也是已知的。同时，视频都是包含音频信息的。YouTube2Text dataset(MSVD dataset)数据集同样由Microsoft Research提供，网址为https://www.microsoft.com/en-us/download/details.aspx?id=52422&from=

http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/default.aspx。该数据集包含1970段YouTube视频片段(时长在10-25s之间),每段视频被标注了大概40条英文句子。

[0116] 实验结果如下表所示。

[0117]

Method	MSVD					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
LSTM-YT [18]	-	-	-	0.333	0.291	-
S2VT [4]	-	-	-	-	0.298	-
TA [17]	0.800	0.647	0.526	0.419	0.296	0.517
TA* [12]	0.811	0.655	0.541	0.422	0.304	0.524
LSTM-E [18]	0.788	0.660	0.554	0.453	0.310	-
HRNE-A [14]	0.792	0.663	0.551	0.438	0.331	-
h-RNN [14]	0.815	0.704	0.604	0.499	0.326	0.658
h-RNN* [14]	0.824	0.711	0.610	0.504	0.329	0.675
SCN [14]	-	-	-	0.511	0.335	0.777
((ML-ATT+(FCIS+VGG)))	0.823	0.715	0.619	0.515	0.341	0.772
((ML-ATT+FCIS))	0.842	0.731	0.633	0.528	0.357	0.786
TM [14]	0.826	0.717	0.619	0.508	0.332	0.694
TM-HQ-SV [14]	0.918	0.872	0.825	0.764	0.429	0.814

[0118] 在这项工作中,本发明提出了一种新的方法来完成视频描述。在各种普及的标准基准上表现出了较好的效果。与以前的工作不同,本发明的方法第一次提出了利用fcis与多层注意力相结合的方法,尽可能的利用视频的有用信息,摒弃无用信息,并模拟现实情况提出动作与声音结合的方式。因此,本发明的方法不仅利用了fcis属性和特征突出化的有点,而且还科学的对不同模态施加注意力,让生成的句子更能够真实的反应视频的内容。本发明的模型能够最大化地提高不同模态信息的利用率。

[0119] 最后,本领域的技术人员能够理解,对本发明的上述实施例能够做出各种修改、变型、以及替换,其均落入如所附权利要求限定的本发明的保护范围。

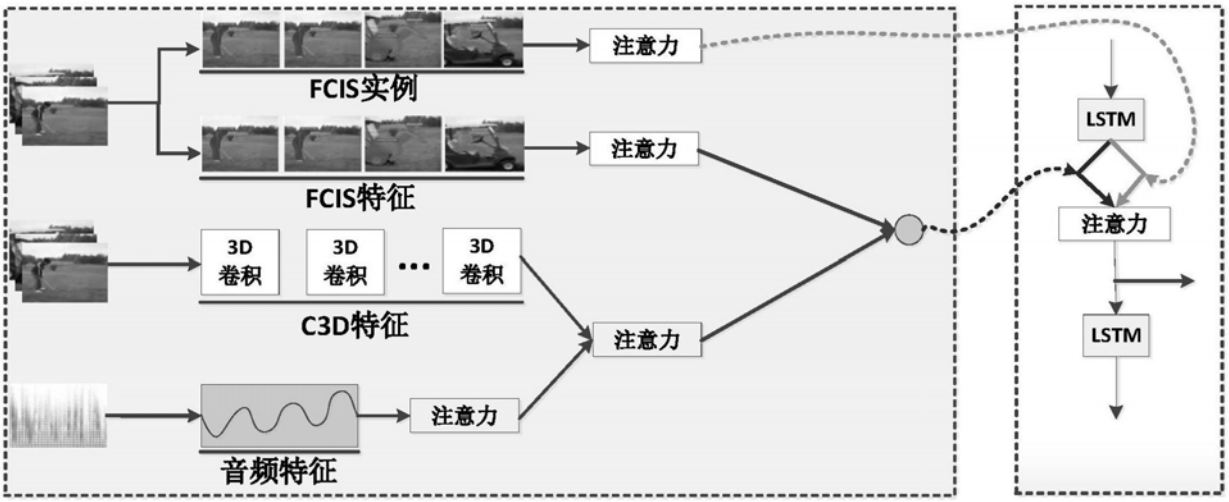

注意力
FCIS实例
注意力
FCIS特征
3D
卷积
3D
卷积
3D
卷积
C3D特征
注意力
注意力
音频特征
LSTM
注意力
LSTM

图1

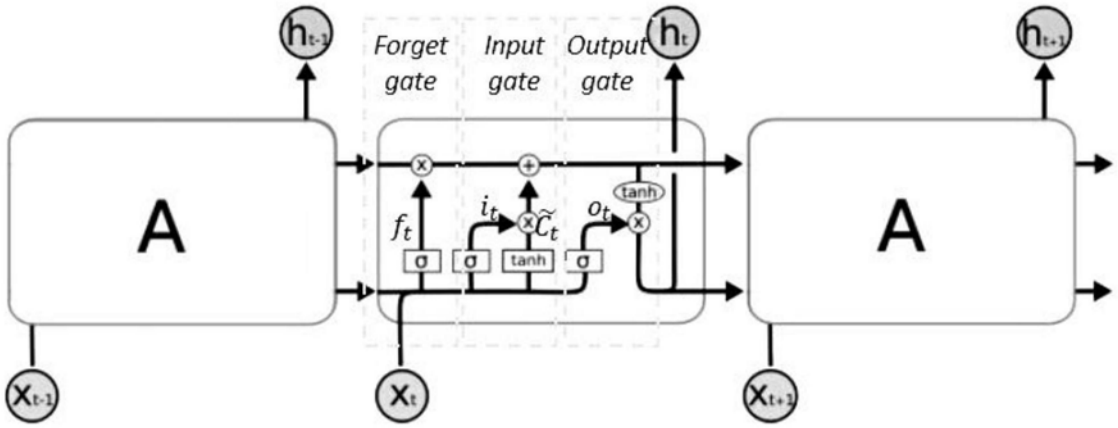

Forget
gate
Input
gate
Output
gate
A
A
tanh
σ
σ
tanh
σ

图2