(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 4 月 23 日 (23.04.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/077858 A1

(51) 国际专利分类号:
G06F 16/73 (2019.01) *G06N 3/04* (2006.01)

(21) 国际申请号: PCT/CN2018/124299

(22) 国际申请日: 2018 年 12 月 27 日 (27.12.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811195508.6 2018年10月15日 (15.10.2018) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY(SHENZHEN)CO.,LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 赵峰 (ZHAO, Feng); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 杜展 (DU, Zhan); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 王健宗 (WANG, Jianzong); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 肖京 (XIAO, Jing); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳中一专利商标事务所 (SHENZHEN ZHONGYI PATENT AND TRADEMARK OFFICE); 中国广东省深圳市福田区深南中路1014号老特区报社四楼（5号信箱）, Guangdong 518028 (CN)。

WO 2020/077858 A1

(54) **Title:** VIDEO DESCRIPTION GENERATION METHOD BASED ON NEURAL NETWORK, AND MEDIUM, TERMINAL AND APPARATUS

(54) **发明名称:** 基于神经网络的视频描述生成方法、介质、终端及装置

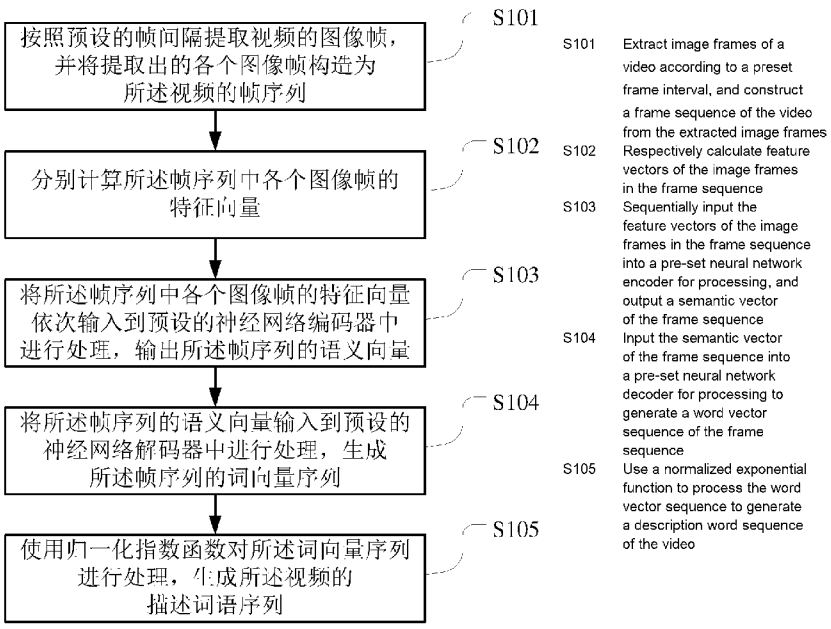

图 1

(57) **Abstract:** The present application belongs to the technical field of computers, and particularly relates to a video description generation method based on a neural network, and a computer-readable storage medium, a terminal device and an apparatus. The method comprises: first extracting image frames of a video according to a pre-set frame interval, and constructing a frame sequence of the video from the extracted image frames; then respectively calculating feature vectors of the image frames in the frame sequence; sequentially inputting the feature vectors of the image frames in the frame sequence into a pre-set neural network encoder for processing, and outputting a semantic vector of the frame sequence; then inputting the semantic vector of the frame sequence into a pre-set neural

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

network decoder for processing to generate a word vector sequence of the frame sequence; and finally using a normalized exponential function to process the word vector sequence to generate a description word sequence of the video. By means of the present application, a neural network is used instead of manual work to carry out video description, without human intervention, thereby greatly improving working efficiency.

(57) 摘要：本申请属于计算机技术领域，尤其涉及一种基于神经网络的视频描述生成方法、计算机可读存储介质、终端设备及装置。所述方法首先按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列，然后分别计算所述帧序列中各个图像帧的特征向量，将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量，再将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列，最后使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。通过本申请，使用神经网络取代人工进行视频描述，无需人力介入，工作效率得到了极大提升。

基于神经网络的视频描述生成方法、介质、终端及装置

本申请要求于 2018 年 10 月 15 日提交中国专利局、申请号为201811195508.6、发明名称为“基于神经网络的视频描述生成方法、存储介质及终端设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请属于计算机技术领域，尤其涉及一种基于神经网络的视频描述生成方法、计算机可读存储介质、终端设备及装置。

背景技术

对视频进行标注、描述的最直接传统的方法是人工的方式，但人工进行视频描述需要耗费大量的时间，效率十分低下，面对目前网络时代的海量视频，此种方式无疑是难以满足实际需要的。

技术问题

有鉴于此，本申请实施例提供了一种基于神经网络的视频描述生成方法、计算机可读存储介质、终端设备及装置，以解决人工进行视频描述需要耗费大量的时间，效率十分低下的问题。

技术解决方案

本申请实施例的第一方面提供了一种基于神经网络的视频描述生成方法，可以包括：

按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列；

分别计算所述帧序列中各个图像帧的特征向量；

将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量；

将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列；

使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。

本申请实施例的第二方面提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令被处理器执行时实现上述视频描述生成方法的步骤。

本申请实施例的第三方面提供了一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现上述视频描述生成方法的步骤。

本申请实施例的第四方面提供了一种视频描述生成装置，可以包括用于实现上述视频描述生成方法的步骤的模块。

有益效果

本申请实施例使用神经网络取代人工进行视频描述，无需人力介入，即可在极短的时间内完成海量的视频描述工作，工作效率得到了极大提升。

附图说明

图 1 为本申请实施例中一种视频描述生成方法的一个实施例流程图；

图 2 为长短期记忆网络中的各个神经网络单元进行数据处理的示意图；

图 3 为本申请实施例中一种视频描述生成装置的一个实施例结构图；

图 4 为本申请实施例中一种终端设备的示意框图。

本发明的实施方式

请参阅图 1，本申请实施例中一种基于神经网络的视频描述生成方法的一个实施例可以包括：

步骤 S101、按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列。

所述帧间隔可以根据实际情况进行设置，例如，可以将其设置为 3 帧、5 帧、10 帧或者其它取值，即每隔 3 帧、5 帧、10 帧或者其它取值从所述视频中提取一个图像帧。

需要注意的是，帧间隔设置的越小，则生成的视频描述结果越准确，但是计算量也越大，会造成较多的系统资源消耗，特别地，当将所述帧间隔设置为 0，即提取所述视频中的所有图像帧，此时生成的视频描述结果最准确，但是计算量也最大，会造成最多的系统资源消耗，反之，帧间隔设置的越大，则生成的视频描述结果越粗略，但是计算量也越小，会造成较少的系统资源消耗，在实际使用中，需要综合考虑系统资源情况以及结果精度要求进行权衡。

在进行完图像帧的提取后，可以根据下式构造所述视频的帧序列：

$$V=\{v_1,v_2,\cdots,v_n,\cdots,v_N\}$$

其中，V 为所述视频的帧序列，n 为从所述视频中提取的图像帧的序号，$1\le n\le N$，N 为从所述视频中提取的图像帧的总数，v_n 为从所述视频中提取的第 n 个图像帧。

步骤 S102、分别计算所述帧序列中各个图像帧的特征向量。

在本实施例中，特征向量的计算在本质上是一种基于局部特征提取的图像表示与相似度度量的方法，局部特征的提取分为两步：提取目标关键点和对关键点进行描述，关键点的定位是目标识别的基础，一般是图像灰度变化的局部极值点，含有显著的结

构性信息，甚至这些点也可以没有实际的直观视觉意义，但却在某种角度、某个尺度上含有丰富的易于匹配的信息。特征点描述即建立特征向量，特征空间的选择决定了图像的哪些特性参与匹配。特征点的特征描述应是对各种变化的不变量，以确保最低限度受位置、视角、尺度和光照等因素的影响。选择合理的特征空间可以降低各类图像变化因素对匹配算法速度、稳健性的影响。

首先，进行基于图像多尺度表示的关键点检测。为了保证提取的特征对于尺度变换保持稳定性能，本实施例在尺度空间内进行图像关键点的检测，尺度空间理论最早出现于计算机视觉领域，其目的是模拟图像数据的多尺度特征。尺度空间理论的主要思想是利用高斯核对原始图像进行尺度变换，以获得多尺度下的图像表示。

关键点检测主要由图像尺度空间生成，高斯差分金字塔建立，候选关键点获取，关键点精定位及其筛选和关键点主方向确定几个部分组成。

图像尺度空间生成主要是对给定的二维图像生成在不同尺度空间下的图像序列图。

高斯差分金字塔建立主要是对尺度空间图像序列进行高斯差分(Difference of Gaussian，DOG)操作，即相邻高斯滤波图像的差分，其主要是为找到具有稳定尺度特征的关键点。

候选关键点的获取主要是在建立的高斯差分空间金字塔中，定义在相邻尺度空间内具有局部最大值或局部最小值的点作为图像关键点的候选。高斯差分尺度空间的中间层的每个像素点与同一层的相邻 8 个像素点、上一层的相邻 9 个像素点和下一层的相邻 9 个像素点总共 26 个相邻像素点进行比较。像素若比相邻的 26 个像素的高斯差分值都大或都小，则该点可作为候选关键点。

关键点精定位及其筛选主要是因为高斯差分图像的像素值对噪声和边缘较敏感，因此，在高斯差分空间检测到的局部极值点要进一步筛选，并重定位到亚像素级和精度尺度位置。同时还要去除低对比度的特征点和不稳定的边缘响应点，以增强匹配稳定性和提高抗噪声能力。

关键点主方向的确定主要目的是保证旋转不变性，基于特征点邻域像素的梯度方向分布特性为每个特征点指定方向参数。在以特征点为中心的邻域窗口内采样，并用梯度方向直方图统计邻域像素的梯度方向。

通过以上步骤，已经完成图像特征点的提取，每个特征点有三个信息：位置、尺度和方向。然后，可以基于梯度方向直方图统计的关键点特征提取。

在图像关键点表示中，仅利用关键点的位置、尺度和方向信息不足以形成足够判决性的特征，于是需要对关键点周围的区域以尺度大小提取灰度统计特征。在特征提

取之前，首先将坐标轴旋转为关键点的方向，以确保旋转不变性。然后以关键点为中心取 8×8 的窗口，在每 4×4 的小块上计算 8 个方向的梯度方向直方图，绘制每个梯度方向的累加值。这种邻域方向性信息联合的思想增强了算法抗噪声的能力，同时对于含有定位误差的特征匹配也提供了较好的容错性。实际计算过程中为增强匹配的稳健性，将扩大特征提取的区域范围，对每个关键点使用 4×4 共 16 个种子点来描述，这样对于一个关键点就可以产生 128 个数据，最终形成 128 维的特征向量。为进一步去除光照对比度变化的影响，可将特征向量的长度归一化。

步骤 S103、将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量。

其中，所述神经网络编码器根据下式进行状态更新：

$$h_t^{(1)} = LSTM^{(1)}(x_t, h_{t-1}^{(1)})$$

其中，$LSTM^{(1)}$ 为预设的第一长短期记忆网络，x_t 为所述第一长短期记忆网络在时刻 t 的输入，$h_t^{(1)}$ 为所述第一长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t\text{-}1}^{(1)}$ 为所述第一长短期记忆网络在时刻 t-1 的隐藏层状态向量。

在本实施例中采用的神经网络编码器是一个长短期记忆网络（Long Short Term Memory，LSTM），LSTM 是循环神经网络的一种，经常用于序列到序列模型。它被广泛应用于机器翻译以及语音识别等领域中。

长短期记忆网络可以通过遗忘门来帮助避免梯度爆炸和消失的问题，当记忆模块中的内容已经过时的时候来重置它们。如图 2 所示，给定 t 时刻的输入 x_t，一个长短期记忆网络中的各个神经网络单元（即神经元）根据下式进行数据处理：

$$\begin{cases} i_t = sigmoid(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t = sigmoid(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t = sigmoid(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \\ c_t = f_t * c_{t-1} + i_t * g_t \\ h_t = o_t * \tanh(c_t) \end{cases}$$

其中，i_t 为输入门在时刻 t 的输出，f_t 为遗忘门在时刻 t 的输出，o_t 为输出门在时刻 t 的输出，W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo}、W_{ho}、W_{xg}、W_{hg}、b_i、b_f、b_o、b_g 均为预设的参数，sigmoid 为 S 型生长函数，且 $sigmoid(x) = \frac{1}{1+e^{-x}}$，tanh 为双曲正切函数，且 $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$。

在该长短期记忆网络中，整个处理过程由三部分组成：即输入门处理、遗忘门处理和输出门处理。

其中，遗忘门是以上一时刻的输出h_{t-1}和本次时刻的输入x_t为输入的 sigmoid 函数，为c_{t-1}中的每一项产生一个在[0,1]内的值，来控制上一时刻的记忆状态被遗忘的程度。输入门和一个 tanh 函数配合控制有哪些新信息被加入。tanh 函数产生一个新的候选向量g_t，输入门为g_t中的每一项产生一个在[0,1]内的值，控制新信息被加入的多少。在有了遗忘门的输出f_t，用来控制上一时刻被遗忘的程度，也有了输入门的输出i_t，用来控制新信息被加入的多少，就可以更新本次时刻的记忆状态了。输出门用来控制本次时刻的记忆状态有多少被过滤掉。先将本次时刻的记忆状态激活，输出门为其中每一项产生一个在[0,1]内的值，控制本次时刻的记忆状态被过滤的程度。

可以看出，通过图 2 所示的结构中的状态结构缓存以往时刻的状态信息，并且通过输入门、遗忘门以及输出门对历史状态信息进行维护，从而实现了长距离历史状态信息的有效缓存。

所述帧序列中各个图像帧的特征向量经过所述神经网络编码器的处理之后，输出为一个固定长度的向量，也即所述帧序列的语义向量。

步骤 S104、将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列。

可选地，所述神经网络解码器可以根据下式进行状态更新：

$$h_t^{(2)} = LSTM^{(2)}([w_{t-1}; h_t^{(1)}], h_{t-1}^{(2)})$$

其中，$LSTM^{(2)}$为预设的第二长短期记忆网络，w_{t-1}为所述第二长短期记忆网络在时刻 t-1 生成的词向量，$[w_{t-1}; h_t^{(1)}]$为由所述第二长短期记忆网络在时刻 t-1 生成的词向量与所述第一长短期记忆网络在时刻 t 的隐藏层状态向量组合成的向量，$h_t^{(2)}$为所述第二长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(2)}$为所述第二长短期记忆网络在时刻 t-1 的隐藏层状态向量。

在本实施例中采用的神经网络解码器也是一个长短期记忆网络，即本实施例中采用了两个 LSTM 网络进行视频描述的生成，分别为作为神经网络编码器的第一长短期记忆网络和作为神经网络解码器的第二长短期记忆网络，所述第二长短期记忆网络与所述第二长短期记忆网络中的各个神经网络单元的数据处理模式类似，具体可参照上述关于第一长短期记忆网络的描述，此处对其不再赘述。

优选地，还可以在此基础上引入时间注意力机制，即所述神经网络解码器可以根据下式进行状态更新：

$$h_t^{(2)} = LSTM^{(2)}([w_{t-1}; h_t^{(1)}], h_{t-1}^{(2)}, c_t)$$

其中，c_t 为所述第二长短期记忆网络在时刻 t 使用的注意力向量。

所述注意力向量可以根据下式进行计算：

$$c_t = \sum_{i=1}^{t} (\alpha_i^{(t)} \times h_i^{(1)})$$

$$= \sum_{i=1}^{t} \left(\frac{e^{f(h_i^{(1)}, h_{t-1}^{(2)})}}{\sum_{i=1}^{t} e^{f(h_i^{(1)}, h_{t-1}^{(2)})}} \times h_i^{(1)}\right)$$

$$\alpha_i^{(t)} = \frac{e^{f(h_i^{(1)}, h_{t-1}^{(2)})}}{\sum_{i=1}^{t} e^{f(h_i^{(1)}, h_{t-1}^{(2)})}}$$

其中，$\alpha_i^{(t)}$ 为预设的权重系数，且 ，$f(h_i^{(1)}, h_{t-1}^{(2)})$ 为预设的对齐函数，可以从以下函数集合中的四个函数中选取一个函数作为所述对齐函数：

$$f(h_i^{(1)}, h_{t-1}^{(2)}) = \begin{cases} h_i^{(1)T} h_{t-1}^{(2)} \\ h_i^{(1)T} W_a h_{t-1}^{(2)} \\ W_a [h_i^{(1)}; h_{t-1}^{(2)}] \\ W_a h_i^{(1)} + W_b h_{t-1}^{(2)} \end{cases}$$

W_a 和 W_b 均为预设的系数。

为了确定哪个对齐函数对模型最合适，可以使用不同的对齐函数，经过梯度下降法得到损失函数收敛后的模型参数，在测试集上生成的视频描述效果中最好的即为最终采用的函数。因为是视频描述生成模型，所以需要综合考量能够评定语句是否符合语音规范的评估分数，以及视频描述与视频内容的相关程度，将这些作为对齐函数的确定基准。

步骤 S105、使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。

在本实施例中，可以根据下式对所述词向量序列进行处理：

$$p(w_t | w_1, w_2, ..., w_{t-1}) = \text{softmax}(W_s h_t^{(2)} + b_s)$$

其中，softmax 为所述归一化指数函数，Ws 和 bs 为预设的系数，$p(w_t | w_1, w_2, ..., w_{t-1})$ 为在时刻 t 的词向量对应的各个词语的概率，这些词语中概率值最大的词语即为时刻 t 的描述词语，将这些描述词语组合在一起，即可生成所述视频的描述词语序列。

综上所述，本申请实施例使用神经网络取代人工进行视频描述，无需人力介入，即可在极短的时间内完成海量的视频描述工作，工作效率得到了极大提升。

对应于上文实施例所述的一种视频描述生成方法，图 3 示出了本申请实施例提供的一种视频描述生成装置的一个实施例结构图。

本实施例中，一种视频描述生成装置可以包括：

帧序列构造模块 301，用于按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列；

特征向量计算模块 302，用于分别计算所述帧序列中各个图像帧的特征向量；

编码器处理模块 303，用于将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量；

解码器处理模块 304，用于将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列；

描述词语序列生成模块 305，用于使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。

进一步地，所述视频描述生成装置还可以包括第一更新模块，用于对所述神经网络编码器根据下式进行状态更新：

$$h_t^{(1)} = LSTM^{(1)}(x_t, h_{t-1}^{(1)})$$

其中，$LSTM^{(1)}$ 为预设的第一长短期记忆网络，x_t 为所述第一长短期记忆网络在时刻 t 的输入，$h_t^{(1)}$ 为所述第一长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(1)}$ 为所述第一长短期记忆网络在时刻 t-1 的隐藏层状态向量；

进一步地，所述视频描述生成装置还可以包括第二更新模块，用于对所述神经网络解码器根据下式进行状态更新：

$$h_t^{(2)} = LSTM^{(2)}([w_{t-1}; h_t^{(1)}], h_{t-1}^{(2)}, c_t)$$

其中，$LSTM^{(2)}$ 为预设的第二长短期记忆网络，w_{t-1} 为所述第二长短期记忆网络在时刻 t-1 生成的词向量，$[w_{t-1}; h_t^{(1)}]$ 为由所述第二长短期记忆网络在时刻 t-1 生成的词向量与所述第一长短期记忆网络在时刻 t 的隐藏层状态向量组合成的向量，c_t 为所述第二长短期记忆网络在时刻 t 使用的注意力向量，$h_t^{(2)}$ 为所述第二长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(2)}$ 为所述第二长短期记忆网络在时刻 t-1 的隐藏层状态向量。

进一步地，所述视频描述生成装置还可以包括注意力向量计算模块，用于根据下式计算所述注意力向量：

$$c_t=\sum_{i=1}^{t}(\alpha_i^{(t)}\times h_i^{(1)})$$
$$=\sum_{i=1}^{t}(\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}\times h_i^{(1)})$$

$$\alpha_i^{(t)}=\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}$$

其中，$\alpha_i^{(t)}$为预设的权重系数，且 ，$f(h_i^{(1)},h_{t-1}^{(2)})$为预设的对齐函数。

进一步地，所述视频描述生成装置还可以包括对齐函数选取模块，用于从以下函数集合中选取一个函数作为所述对齐函数：

$$f(h_i^{(1)},h_{t-1}^{(2)})=\begin{cases}h_i^{(1)T}h_{t-1}^{(2)}\\ h_i^{(1)T}W_a h_{t-1}^{(2)}\\ W_a[h_i^{(1)};h_{t-1}^{(2)}]\\ W_a h_i^{(1)}+W_b h_{t-1}^{(2)}\end{cases}$$

其中，W_a和W_b均为预设的系数。

进一步地，所述视频描述生成装置还可以包括数据处理模块，用于所述第一长短期记忆网络和所述第二长短期记忆网络中的各个神经网络单元根据下式进行数据处理：

$$\begin{cases}i_t=sigmoid(W_{xi}x_t+W_{hi}h_{t-1}+b_i)\\ f_t=sigmoid(W_{xf}x_t+W_{hf}h_{t-1}+b_f)\\ o_t=sigmoid(W_{xo}x_t+W_{ho}h_{t-1}+b_o)\\ g_t=\tanh(W_{xg}x_t+W_{hg}h_{t-1}+b_g)\\ c_t=f_t*c_{t-1}+i_t*g_t\\ h_t=o_t*\tanh(c_t)\end{cases}$$

其中，i_t为输入门在时刻 t 的输出，f_t为遗忘门在时刻 t 的输出，o_t为输出门在时刻 t 的输出，W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo}、W_{ho}、W_{xg}、W_{hg}、b_i、b_f、b_o、b_g均为预设的参数，sigmoid 为 S 型生长函数，且$sigmoid(x)=\frac{1}{1+e^{-x}}$，tanh 为双曲正切函数，且$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置，

模块和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

图 4 示出了本申请实施例提供的一种终端设备的示意框图，为了便于说明，仅示出了与本申请实施例相关的部分。

在本实施例中，所述终端设备 4 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。该终端设备 4 可包括：处理器 40、存储器 41 以及存储在所述存储器 41 中并可在所述处理器 40 上运行的计算机可读指令 42，例如执行上述的视频描述生成方法的计算机可读指令。所述处理器 40 执行所述计算机可读指令 42 时实现上述各个视频描述生成方法实施例中的步骤。

在本申请各个实施例中的各功能单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干计算机可读指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、磁碟或者光盘等各种可以存储计算机可读指令的介质。

权利要求书

1. 一种基于神经网络的视频描述生成方法，其特征在于，包括：

按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列；

分别计算所述帧序列中各个图像帧的特征向量；

将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量；

将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列；

使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。

2. 根据权利要求 1 所述的视频描述生成方法，其特征在于，所述神经网络编码器根据下式进行状态更新：

$$h_t^{(1)} = LSTM^{(1)}(x_t, h_{t-1}^{(1)})$$

其中，$LSTM^{(1)}$ 为预设的第一长短期记忆网络，x_t 为所述第一长短期记忆网络在时刻 t 的输入，$h_t^{(1)}$ 为所述第一长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(1)}$ 为所述第一长短期记忆网络在时刻 t-1 的隐藏层状态向量；

所述神经网络解码器根据下式进行状态更新：

$$h_t^{(2)} = LSTM^{(2)}([w_{t-1}; h_t^{(1)}], h_{t-1}^{(2)}, c_t)$$

其中，$LSTM^{(2)}$ 为预设的第二长短期记忆网络，w_{t-1} 为所述第二长短期记忆网络在时刻 t-1 生成的词向量，$[w_{t-1}; h_t^{(1)}]$ 为由所述第二长短期记忆网络在时刻 t-1 生成的词向量与所述第一长短期记忆网络在时刻 t 的隐藏层状态向量组合成的向量，c_t 为所述第二长短期记忆网络在时刻 t 使用的注意力向量，$h_t^{(2)}$ 为所述第二长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(2)}$ 为所述第二长短期记忆网络在时刻 t-1 的隐藏层状态向量。

3. 根据权利要求 2 所述的视频描述生成方法，其特征在于，所述注意力向量的设置过程包括：

根据下式计算所述注意力向量：

$$c_t=\sum_{i=1}^{t}(\alpha_i^{(t)}\times h_i^{(1)})$$
$$=\sum_{i=1}^{t}(\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}\times h_i^{(1)})$$

$$\alpha_i^{(t)}=\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}$$

其中，$\alpha_i^{(t)}$为预设的权重系数，且 ，$f(h_i^{(1)},h_{t-1}^{(2)})$为预设的对齐函数。

4. 根据权利要求 3 所述的视频描述生成方法，其特征在于，所述对齐函数的设置过程包括：

从以下函数集合中选取一个函数作为所述对齐函数：

$$f(h_i^{(1)},h_{t-1}^{(2)})=\begin{cases}h_i^{(1)T}h_{t-1}^{(2)}\\ h_i^{(1)T}W_a h_{t-1}^{(2)}\\ W_a[h_i^{(1)};h_{t-1}^{(2)}]\\ W_a h_i^{(1)}+W_b h_{t-1}^{(2)}\end{cases}$$

其中，W_a和W_b均为预设的系数。

5. 根据权利要求 2 至 4 中任一项所述的视频描述生成方法，其特征在于，所述第一长短期记忆网络和所述第二长短期记忆网络中的各个神经网络单元根据下式进行数据处理：

$$\begin{cases}i_t=sigmoid(W_{xi}x_t+W_{hi}h_{t-1}+b_i)\\ f_t=sigmoid(W_{xf}x_t+W_{hf}h_{t-1}+b_f)\\ o_t=sigmoid(W_{xo}x_t+W_{ho}h_{t-1}+b_o)\\ g_t=\tanh(W_{xg}x_t+W_{hg}h_{t-1}+b_g)\\ c_t=f_t*c_{t-1}+i_t*g_t\\ h_t=o_t*\tanh(c_t)\end{cases}$$

其中，i_t为输入门在时刻 t 的输出，f_t为遗忘门在时刻 t 的输出，o_t为输出门在时刻 t 的输出，W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo}、W_{ho}、W_{xg}、W_{hg}、b_i、b_f、b_o、b_g均为预设的参数，sigmoid 为 S 型生长函数，且$sigmoid(x)=\frac{1}{1+e^{-x}}$，tanh 为双曲正切函数，且$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$。

6. 一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，其特征在于，所述计算机可读指令被处理器执行时实现如下步骤：

按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列；

分别计算所述帧序列中各个图像帧的特征向量；

将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量；

将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列；

使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。

7. 根据权利要求 6 所述的计算机可读存储介质，其特征在于，所述神经网络编码器根据下式进行状态更新：

$$h_t^{(1)} = LSTM^{(1)}(x_t, h_{t-1}^{(1)})$$

其中，$LSTM^{(1)}$ 为预设的第一长短期记忆网络，x_t 为所述第一长短期记忆网络在时刻 t 的输入，$h_t^{(1)}$ 为所述第一长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(1)}$ 为所述第一长短期记忆网络在时刻 t-1 的隐藏层状态向量；

所述神经网络解码器根据下式进行状态更新：

$$h_t^{(2)} = LSTM^{(2)}([w_{t-1}; h_t^{(1)}], h_{t-1}^{(2)}, c_t)$$

其中，$LSTM^{(2)}$ 为预设的第二长短期记忆网络，w_{t-1} 为所述第二长短期记忆网络在时刻 t-1 生成的词向量，$[w_{t-1}; h_t^{(1)}]$ 为由所述第二长短期记忆网络在时刻 t-1 生成的词向量与所述第一长短期记忆网络在时刻 t 的隐藏层状态向量组合成的向量，c_t 为所述第二长短期记忆网络在时刻 t 使用的注意力向量，$h_t^{(2)}$ 为所述第二长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(2)}$ 为所述第二长短期记忆网络在时刻 t-1 的隐藏层状态向量。

8. 根据权利要求 7 所述的计算机可读存储介质，其特征在于，所述注意力向量的设置过程包括：

根据下式计算所述注意力向量：

$$c_t=\sum_{i=1}^{t}(\alpha_i^{(t)}\times h_i^{(1)})$$
$$=\sum_{i=1}^{t}(\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}\times h_i^{(1)})$$

$$\alpha_i^{(t)}=\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}$$

其中，$\alpha_i^{(t)}$为预设的权重系数，且 ，$f(h_i^{(1)},h_{t-1}^{(2)})$为预设的对齐函数。

9. 根据权利要求 8 所述的计算机可读存储介质，其特征在于，所述对齐函数的设置过程包括：

从以下函数集合中选取一个函数作为所述对齐函数：

$$f(h_i^{(1)},h_{t-1}^{(2)})=\begin{cases}h_i^{(1)T}h_{t-1}^{(2)}\\ h_i^{(1)T}W_a h_{t-1}^{(2)}\\ W_a[h_i^{(1)};h_{t-1}^{(2)}]\\ W_a h_i^{(1)}+W_b h_{t-1}^{(2)}\end{cases}$$

其中，W_a和W_b均为预设的系数。

10. 根据权利要求 7 至 9 中任一项所述的计算机可读存储介质，其特征在于，所述第一长短期记忆网络和所述第二长短期记忆网络中的各个神经网络单元根据下式进行数据处理：

$$\begin{cases}i_t=sigmoid(W_{xi}x_t+W_{hi}h_{t-1}+b_i)\\ f_t=sigmoid(W_{xf}x_t+W_{hf}h_{t-1}+b_f)\\ o_t=sigmoid(W_{xo}x_t+W_{ho}h_{t-1}+b_o)\\ g_t=\tanh(W_{xg}x_t+W_{hg}h_{t-1}+b_g)\\ c_t=f_t*c_{t-1}+i_t*g_t\\ h_t=o_t*\tanh(c_t)\end{cases}$$

其中，i_t为输入门在时刻 t 的输出，f_t为遗忘门在时刻 t 的输出，o_t为输出门在时刻 t 的输出，W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo}、W_{ho}、W_{xg}、W_{hg}、b_i、b_f、b_o、b_g均为预设的参数，sigmoid 为 S 型生长函数，且$sigmoid(x)=\frac{1}{1+e^{-x}}$，tanh 为双曲正切函数，且$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$。

11. 一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，其特征在于，所述处理器执行所述计算机可读指令时实现如下步骤：

按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列；

分别计算所述帧序列中各个图像帧的特征向量；

将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量；

将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列；

使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。

12. 根据权利要求 11 所述的终端设备，其特征在于，所述神经网络编码器根据下式进行状态更新：

$$h_t^{(1)} = LSTM^{(1)}(x_t, h_{t-1}^{(1)})$$

其中，$LSTM^{(1)}$ 为预设的第一长短期记忆网络，x_t 为所述第一长短期记忆网络在时刻 t 的输入，$h_t^{(1)}$ 为所述第一长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(1)}$ 为所述第一长短期记忆网络在时刻 t-1 的隐藏层状态向量；

所述神经网络解码器根据下式进行状态更新：

$$h_t^{(2)} = LSTM^{(2)}([w_{t-1}; h_t^{(1)}], h_{t-1}^{(2)}, c_t)$$

其中，$LSTM^{(2)}$ 为预设的第二长短期记忆网络，w_{t-1} 为所述第二长短期记忆网络在时刻 t-1 生成的词向量，$[w_{t-1}; h_t^{(1)}]$ 为由所述第二长短期记忆网络在时刻 t-1 生成的词向量与所述第一长短期记忆网络在时刻 t 的隐藏层状态向量组合成的向量，c_t 为所述第二长短期记忆网络在时刻 t 使用的注意力向量，$h_t^{(2)}$ 为所述第二长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(2)}$ 为所述第二长短期记忆网络在时刻 t-1 的隐藏层状态向量。

13. 根据权利要求 12 所述的终端设备，其特征在于，所述注意力向量的设置过程包括：

根据下式计算所述注意力向量：

$$c_t=\sum_{i=1}^{t}(\alpha_i^{(t)}\times h_i^{(1)})$$
$$=\sum_{i=1}^{t}(\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}\times h_i^{(1)})$$

$$\alpha_i^{(t)}=\frac{e^{f(h_i^{(1)},h_{t-1}^{(2)})}}{\sum_{i=1}^{t}e^{f(h_i^{(1)},h_{t-1}^{(2)})}}$$

其中，$\alpha_i^{(t)}$为预设的权重系数，且 ，$f(h_i^{(1)},h_{t-1}^{(2)})$为预设的对齐函数。

14. 根据权利要求 13 所述的终端设备，其特征在于，所述对齐函数的设置过程包括：

从以下函数集合中选取一个函数作为所述对齐函数：

$$f(h_i^{(1)},h_{t-1}^{(2)})=\begin{cases}h_i^{(1)T}h_{t-1}^{(2)}\\ h_i^{(1)T}W_a h_{t-1}^{(2)}\\ W_a[h_i^{(1)};h_{t-1}^{(2)}]\\ W_a h_i^{(1)}+W_b h_{t-1}^{(2)}\end{cases}$$

其中，W_a和W_b均为预设的系数。

15. 根据权利要求 12 至 14 中任一项所述的终端设备，其特征在于，所述第一长短期记忆网络和所述第二长短期记忆网络中的各个神经网络单元根据下式进行数据处理：

$$\begin{cases}i_t=sigmoid(W_{xi}x_t+W_{hi}h_{t-1}+b_i)\\ f_t=sigmoid(W_{xf}x_t+W_{hf}h_{t-1}+b_f)\\ o_t=sigmoid(W_{xo}x_t+W_{ho}h_{t-1}+b_o)\\ g_t=\tanh(W_{xg}x_t+W_{hg}h_{t-1}+b_g)\\ c_t=f_t*c_{t-1}+i_t*g_t\\ h_t=o_t*\tanh(c_t)\end{cases}$$

其中，i_t为输入门在时刻 t 的输出，f_t为遗忘门在时刻 t 的输出，o_t为输出门在时刻 t 的输出，W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo}、W_{ho}、W_{xg}、W_{hg}、b_i、b_f、b_o、b_g均为预设的参数，sigmoid 为 S 型生长函数，且$sigmoid(x)=\frac{1}{1+e^{-x}}$，tanh 为双曲正切函数，且$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$。

16、一种视频描述生成装置，其特征在于，包括：

帧序列构造模块，用于按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列；

特征向量计算模块，用于分别计算所述帧序列中各个图像帧的特征向量；

编码器处理模块，用于将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量；

解码器处理模块，用于将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列；

描述词语序列生成模块，用于使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列。

17. 根据权利要求 16 所述的视频描述生成装置，其特征在于，还包括：

第一更新模块，用于对所述神经网络编码器根据下式进行状态更新：

$$h_t^{(1)} = LSTM^{(1)}(x_t, h_{t-1}^{(1)})$$

其中，$LSTM^{(1)}$ 为预设的第一长短期记忆网络，x_t 为所述第一长短期记忆网络在时刻 t 的输入，$h_t^{(1)}$ 为所述第一长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(1)}$ 为所述第一长短期记忆网络在时刻 t-1 的隐藏层状态向量；

第二更新模块，用于对所述神经网络解码器根据下式进行状态更新：

$$h_t^{(2)} = LSTM^{(2)}([w_{t-1}; h_t^{(1)}], h_{t-1}^{(2)}, c_t)$$

其中，$LSTM^{(2)}$ 为预设的第二长短期记忆网络，w_{t-1} 为所述第二长短期记忆网络在时刻 t-1 生成的词向量，$[w_{t-1}; h_t^{(1)}]$ 为由所述第二长短期记忆网络在时刻 t-1 生成的词向量与所述第一长短期记忆网络在时刻 t 的隐藏层状态向量组合成的向量，c_t 为所述第二长短期记忆网络在时刻 t 使用的注意力向量，$h_t^{(2)}$ 为所述第二长短期记忆网络在时刻 t 的隐藏层状态向量，$h_{t-1}^{(2)}$ 为所述第二长短期记忆网络在时刻 t-1 的隐藏层状态向量。

18. 根据权利要求 17 所述的视频描述生成装置，其特征在于，还包括：

注意力向量计算模块，用于根据下式计算所述注意力向量：

$$c_t = \sum_{i=1}^{t} (\alpha_i^{(t)} \times h_i^{(1)})$$

$$= \sum_{i=1}^{t} \left(\frac{e^{f(h_i^{(1)}, h_{t-1}^{(2)})}}{\sum_{i=1}^{t} e^{f(h_i^{(1)}, h_{t-1}^{(2)})}} \times h_i^{(1)} \right)$$

$$\alpha_i^{(t)} = \frac{e^{f(h_i^{(1)}, h_{t-1}^{(2)})}}{\sum_{i=1}^{t} e^{f(h_i^{(1)}, h_{t-1}^{(2)})}}$$

其中，$\alpha_i^{(t)}$ 为预设的权重系数，且 ，$f(h_i^{(1)}, h_{t-1}^{(2)})$ 为预设的对齐函数。

19. 根据权利要求 18 所述的视频描述生成装置，其特征在于，还包括：

对齐函数选取模块，用于从以下函数集合中选取一个函数作为所述对齐函数：

$$f(h_i^{(1)}, h_{t-1}^{(2)}) = \begin{cases} h_i^{(1)T} h_{t-1}^{(2)} \\ h_i^{(1)T} W_a h_{t-1}^{(2)} \\ W_a [h_i^{(1)}; h_{t-1}^{(2)}] \\ W_a h_i^{(1)} + W_b h_{t-1}^{(2)} \end{cases}$$

其中，W_a 和 W_b 均为预设的系数。

20. 根据权利要求 17 至 19 中任一项所述的视频描述生成装置，其特征在于，还包括：

数据处理模块，用于所述第一长短期记忆网络和所述第二长短期记忆网络中的各个神经网络单元根据下式进行数据处理：

$$\begin{cases} i_t = sigmoid(W_{xi} x_t + W_{hi} h_{t-1} + b_i) \\ f_t = sigmoid(W_{xf} x_t + W_{hf} h_{t-1} + b_f) \\ o_t = sigmoid(W_{xo} x_t + W_{ho} h_{t-1} + b_o) \\ g_t = \tanh(W_{xg} x_t + W_{hg} h_{t-1} + b_g) \\ c_t = f_t * c_{t-1} + i_t * g_t \\ h_t = o_t * \tanh(c_t) \end{cases}$$

其中，i_t 为输入门在时刻 t 的输出，f_t 为遗忘门在时刻 t 的输出，o_t 为输出门在时刻 t 的输出，W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo}、W_{ho}、W_{xg}、W_{hg}、b_i、b_f、b_o、b_g 均为预设的参数，sigmoid 为 S 型生长函数，且 $sigmoid(x) = \frac{1}{1+e^{-x}}$，tanh 为双曲正切函数，且 $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$。

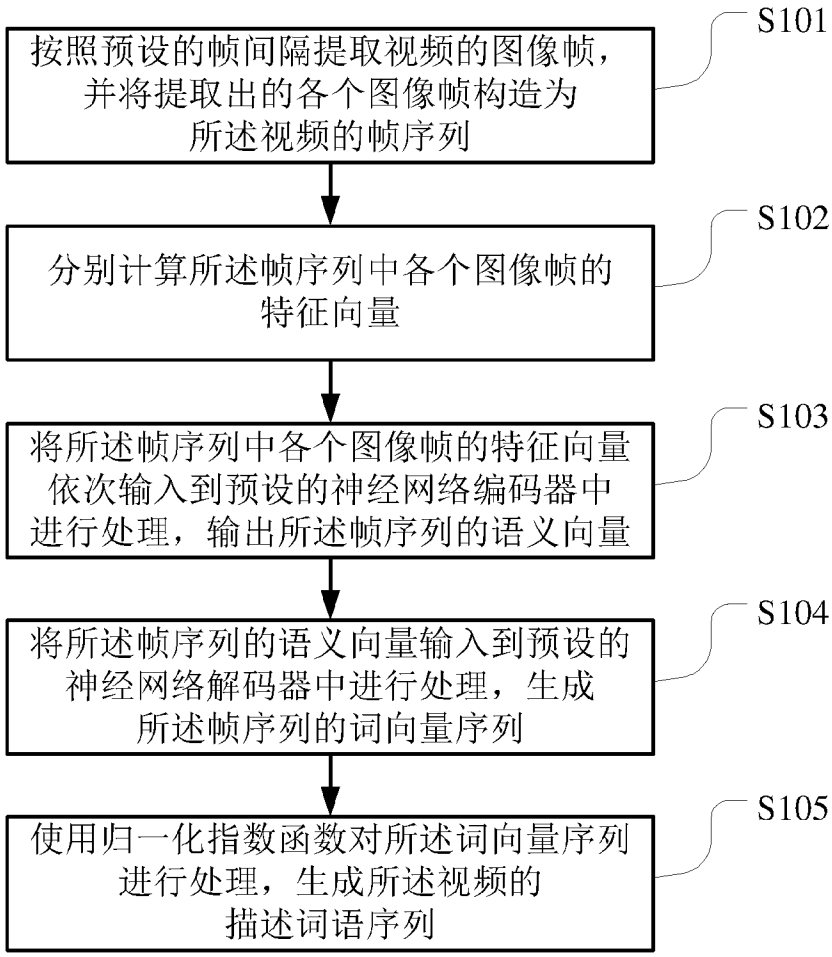
S101
按照预设的帧间隔提取视频的图像帧，并将提取出的各个图像帧构造为所述视频的帧序列
S102
分别计算所述帧序列中各个图像帧的特征向量
S103
将所述帧序列中各个图像帧的特征向量依次输入到预设的神经网络编码器中进行处理，输出所述帧序列的语义向量
S104
将所述帧序列的语义向量输入到预设的神经网络解码器中进行处理，生成所述帧序列的词向量序列
S105
使用归一化指数函数对所述词向量序列进行处理，生成所述视频的描述词语序列

图 1

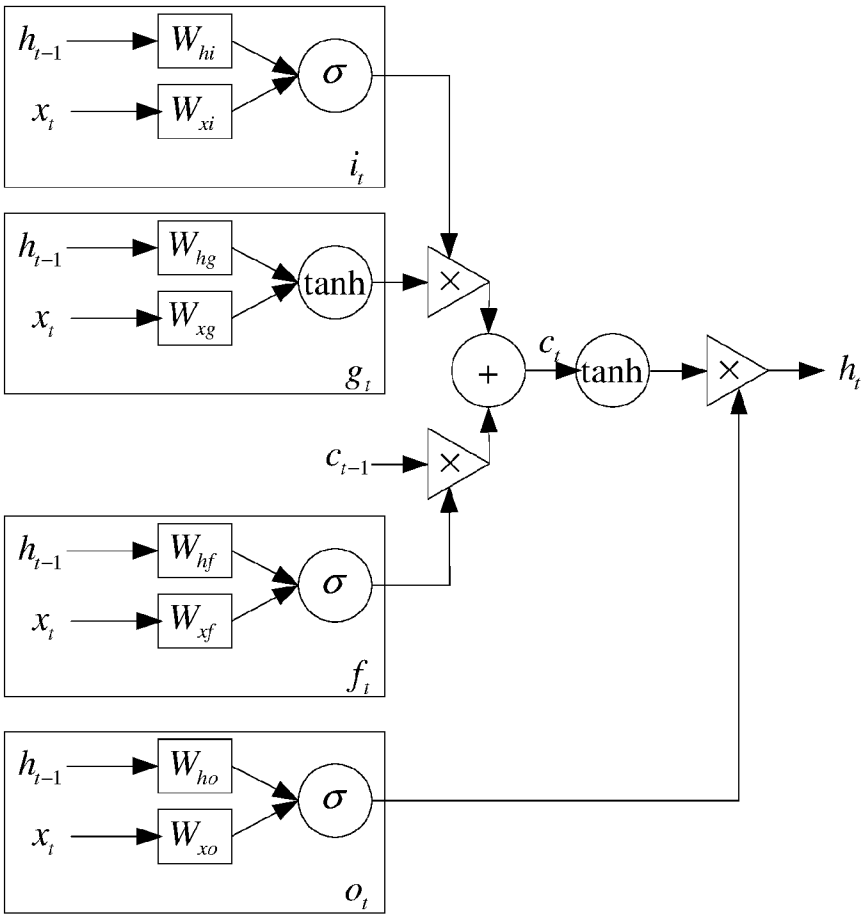
h_{t-1}
W_{hi}
x_t
W_{xi}
σ
i_t
h_{t-1}
W_{hg}
x_t
W_{xg}
tanh
g_t
×
+
c_t
tanh
×
h_t
c_{t-1}
×
h_{t-1}
W_{hf}
x_t
W_{xf}
σ
f_t
h_{t-1}
W_{ho}
x_t
W_{xo}
σ
o_t

图 2

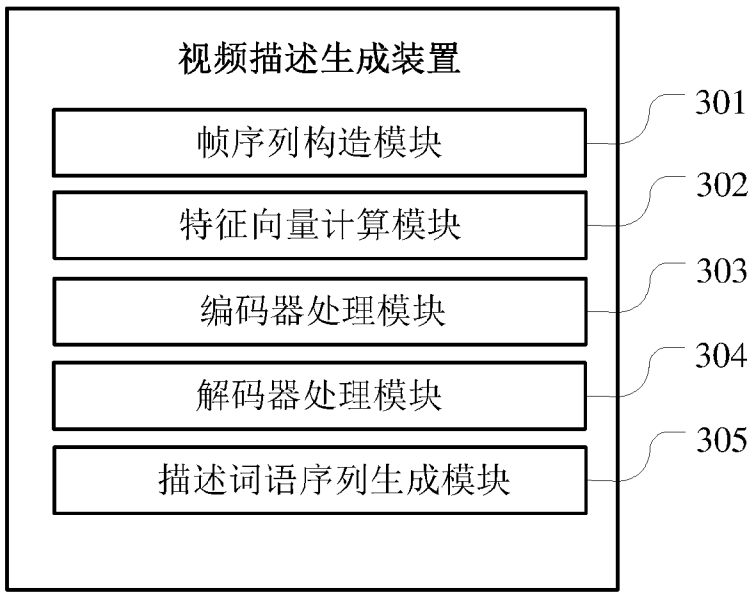
视频描述生成装置
帧序列构造模块
特征向量计算模块
编码器处理模块
解码器处理模块
描述词语序列生成模块
301
302
303
304
305

图 3

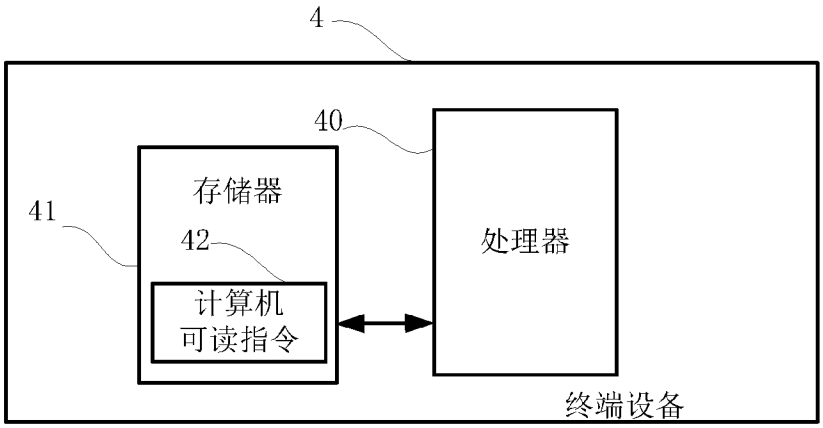

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/124299

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/73(2019.01)i; G06N 3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F,G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT; CNABS; CNKI; USTXT; VEN; BAIDU; GOOGLE: 视频, 描述, 标注, 向量, 长短时记忆网络, 长短期记忆网络, video, description, vector, long short-term memory network

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	VENUGOPALAN, S. et al. "Sequence to Sequence - Video to Text" *2015 IEEE International Conference on Computer Vision (ICCV)*, 19 October 2015 (2015-10-19), pages 4534-4542	1, 2, 4-7, 9-12, 14-17, 19, 20
Y	VENUGOPALAN, S. et al. "Sequence to Sequence - Video to Text" *2015 IEEE International Conference on Computer Vision (ICCV)*, 19 October 2015 (2015-10-19), pages 4534-4542	3, 8, 13, 18
Y	CN 108416065 A (FUDAN UNIVERSITY) 17 August 2018 (2018-08-17) description, paragraphs 28-60	3, 8, 13, 18
X	CN 108388900 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 10 August 2018 (2018-08-10) description, paragraphs 83-142	1-20
X	CN 107391646 A (TSINGHUA UNIVERSITY) 24 November 2017 (2017-11-24) description, paragraphs 113-227	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 June 2019	**04 July 2019**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/124299

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 108416065 A	17 August 2018	None	
CN 108388900 A	10 August 2018	None	
CN 107391646 A	24 November 2017	None	

国际检索报告

国际申请号
PCT/CN2018/124299

A. 主题的分类

G06F 16/73(2019.01)i; G06N 3/04(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F,G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNTXT;CNABS;CNKI;USTXT;VEN;BAIDU;GOOGLE:视频，描述，标注，向量，长短时记忆网络，长短期记忆网络，video, description, vector, long short-term memory network

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	Subhashini Venugopalan等. "Sequence to Sequence - Video to Text" 2015 IEEE International Conference on Computer Vision (ICCV), 2015年 10月 19日 (2015-10-19), 第4534-4542页	1，2，4-7，9-12，14-17，19，20
Y	Subhashini Venugopalan等. "Sequence to Sequence - Video to Text" 2015 IEEE International Conference on Computer Vision (ICCV), 2015年 10月 19日 (2015-10-19), 第4534-4542页	3，8，13，18
Y	CN 108416065 A (复旦大学) 2018年 8月 17日 (2018-08-17) 说明书第28-60段	3，8，13，18
X	CN 108388900 A (华南理工大学) 2018年 8月 10日 (2018-08-10) 说明书第83-142段	1-20
X	CN 107391646 A (清华大学) 2017年 11月 24日 (2017-11-24) 说明书第113-227段	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2019年 6月 26日	2019年 7月 4日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 高巍巍 电话号码 86-（010）-62089149

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/124299

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 108416065 A	2018年 8月 17日	无	
CN 108388900 A	2018年 8月 10日	无	
CN 107391646 A	2017年 11月 24日	无	