(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 109508400 B
(45)授权公告日 2020.08.28

(21)申请号 201811172666.X

(22)申请日 2018.10.09

(65)同一申请的已公布的文献号
申请公布号 CN 109508400 A

(43)申请公布日 2019.03.22

(73)专利权人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路95号

(72)发明人 周玉 朱军楠 张家俊 宗成庆

(74)专利代理机构 北京市恒有知识产权代理事务所(普通合伙) 11576
代理人 郭文浩

(51)Int.Cl.
G06F 16/583(2019.01)
G06F 16/36(2019.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)

(56)对比文件
CN 107608943 A,2018.01.19
CN 107918782 A,2018.04.17

审查员 刘洋

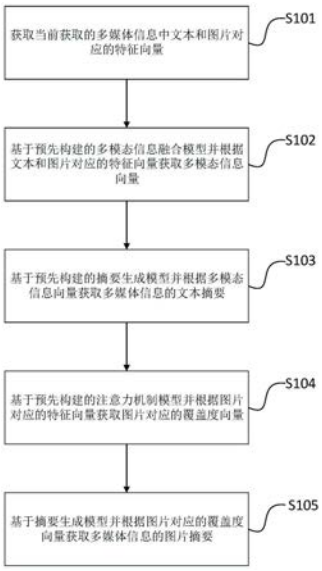
权利要求书3页 说明书8页 附图3页

(54)发明名称

图文摘要生成方法

(57)摘要

本发明属于自然语言技术领域,具体提供了一种图文摘要生成方法,旨在解决现有技术图片和文本不对齐导致摘要信息不准确的问题。为此目的,本发明提供了一种图文摘要生成方法,包括获取多媒体信息中文本和图片对应的特征向量;根据文本和图片对应的特征向量获取多模态信息向量;基于预先构建的摘要生成模型并根据多模态信息向量获取多媒体信息的文本摘要;根据图片对应的特征向量获取图片对应的覆盖度向量;基于摘要生成模型并根据图片对应的覆盖度向量获取多媒体信息的图片摘要;将文本摘要和图片摘要结合作为多媒体信息的图文摘要。基于上述步骤,本发明提供的方法可以得到更准确表现多媒体信息内容的图文摘要。



CN 109508400 B

1.一种图文摘要生成方法,其特征在于,包括:

获取当前获取的多媒体信息中文本和图片对应的特征向量;

基于预先构建的多模态信息融合模型并根据所述文本和图片对应的特征向量获取多模态信息向量;

基于预先构建的摘要生成模型并根据所述多模态信息向量获取所述多媒体信息的文本摘要;

基于预先构建的注意力机制模型并根据图片对应的特征向量获取所述图片对应的覆盖度向量;

基于所述摘要生成模型并根据所述图片对应的覆盖度向量获取所述多媒体信息的图片摘要;

将所述文本摘要和图片摘要结合作为所述多媒体信息的图文摘要;

其中,所述多模态信息融合模型、摘要生成模型以及注意力机制模型均是基于预设的多媒体信息训练数据集并利用机器学习算法所构建的神经网络模型。

2.根据权利要求1所述的图文摘要生成方法,其特征在于,“获取当前获取的多媒体信息中文本和图片对应的特征向量”的步骤包括:

根据下式所示的双向长短期记忆网络获取所述多媒体信息中文本的特征向量:

$f_t=\sigma_g(W_f x_t+U_f c_{t-1}+b_f)$

$i_t=\sigma_g(W_i x_t+U_i c_{t-1}+b_i)$

$o_t=\sigma_g(W_o x_t+U_o C_{t-1}+b_o)$

$c_t=f_t\odot c_{t-1}+i_t\odot\sigma_c(W_c x_t+U_c h_{t-1}+b_c)$

$h_t=o_t\odot\sigma_h(c_t)$

其中,f_t、i_t、o_t分别表示所述双向长短期记忆网络在t时刻的遗忘门、输入门和输出门的输出,σ_g、σ_c、σ_h分别表示遗忘门、输入门和输出门的激活函数,W_f、W_i、W_o分别表示遗忘门、输入门和输出门的第一矩阵参数,U_f、U_i、U_o分别表示遗忘门、输入门和输出门的第二矩阵参数,x_t表示在t时刻的输入的文本词向量,c_{t-1}表示在t-1时刻的文本的特征向量,b_f、b_i、b_o分别表示遗忘门、输入门和输出门的偏置参数,h_t表示t时刻文本的特征向量对应的隐层向量;

基于预先构建的图片特征提取模型获取所述多媒体信息中图片的fc7特征或者pool5特征,将所述fc7特征或者pool5特征转换为图片对应的特征向量;

其中,所述图片特征提取模型是基于预设的图片数据集并利用机器学习算法所构建的神经网络模型。

3.根据权利要求2所述的图文摘要生成方法,其特征在于,“将所述fc7特征或者pool5特征转换为图片对应的特征向量”的步骤包括:

将所述fc7特征与预先获取的图片的特征向量的注意力分布相乘,得到所述图片对应的特征向量;或者

将所述pool5特征与预先获取的图片的特征向量的注意力分布相乘,得到所述图片对应的特征向量;或者

获取图片多个区域的注意力分布,根据所述图片多个区域的注意力分布以及图片多个区域对应的向量进行加权求和,将加权求和的结果与预先获取的图片的特征向量的注意力

分布相乘，得到所述图片对应的特征向量。

4.根据权利要求1所述的图文摘要生成方法，其特征在于，“基于预先构建的多模态信息融合模型并根据文本的特征向量和图片的特征向量获取多模态信息向量”的步骤包括：

根据下式所述的注意力机制获取所述多模态信息向量：

$$\alpha_{txt}^{t} = \sigma(W_{txt}c_{txt}^{t} + U_{txt}s_t)$$

$$\alpha_{img}^{t} = \sigma(W_{img}c_{img}^{t} + U_{img}s_t)$$

$$c_{mm}^{t} = \alpha_{txt}^{t}c_{txt}^{t} + \alpha_{img}^{t}c_{img}^{t}$$

其中，α_{txt}^{t}、α_{img}^{t}分别表示文本和图片的特征向量的注意力分布，σ表示激活函数，W_{txt}、W_{img}分别表示所述多模态信息融合模型的第一矩阵参数，c_{txt}^{t}、c_{img}^{t}分别表示文本和图片的特征向量，U_{txt}、U_{img}分别表示所述多模态信息融合模型的第二矩阵参数，s_t表示所述多模态信息融合模型的状态参数，c_{mm}^{t}表示所述多模态信息向量。

5.根据权利要求1所述的图文摘要生成方法，其特征在于，在“基于预先构建的摘要生成模型并根据所述多模态信息向量获取所述多媒体信息的文本摘要”的步骤之前，所述方法还包括：

基于预先获取的多模态信息向量并利用注意力机制计算从预设的历史词库中生成和/或复制所述多模态信息中文本的概率；

根据所述概率并利用负对数似然损失函数以及覆盖度损失函数优化所述摘要生成模型的参数。

6.根据权利要求5所述的图文摘要生成方法，其特征在于，“根据所述概率并利用负对数似然损失函数以及覆盖度损失函数优化所述摘要生成模型的参数”的步骤包括：

按照下式所示的方法优化所述摘要生成模型的参数：

$$p_g = \sigma(W_h^{*}c_{mm} + W_s^{*}s_t + W_x x_t)$$

$$p_w = p_g p_v(w) + (1 - p_g)\sum_{w_i = w}\alpha_i^{t}$$

$$L_t = -\log(p_{w_t}) + \sum_{i}\min(\alpha_i^{t}, cov_i^{t})$$

其中，p_g表示从预设的历史词库中生成词的概率，σ表示激活函数，W_h^{*}、W_s^{*}、W_X均表示摘要生成模型的矩阵参数，c_{mm}表示多模态信息向量，s_t表示摘要生成模型的状态参数，p_w表示一个词生成和/或复制的概率，$p_v(w)$表示从预设的历史词库中生成词w的概率，α_i^{t}表示t时刻第i个词的文本注意力分布，L_t表示负对数似然损失和覆盖度损失，p_{w_t}表示t时刻从预设的历史词库中生成词或者从输入文本中复制词的概率分布，cov_i^{t}表示t时刻第i个词的文本覆盖度向量，x_t表示在t时刻的输入的文本词向量。

7.根据权利要求1所述的图文摘要生成方法，其特征在于，“基于预先构建的注意力机制模型并根据图片对应的特征向量获取所述图片对应的覆盖度向量”的步骤包括：

基于所述注意力机制模型获取所述图片对应的特征向量多个时刻的注意力分布，将所述多个时刻的注意力分布累加得到所述图片对应的覆盖度向量。

8.根据权利要求1所述的图文摘要生成方法，其特征在于，“基于所述摘要生成模型并根据所述图片对应的覆盖度向量获取所述多媒体信息的图片摘要”的步骤包括：

基于所述摘要生成模型获取每张图片的覆盖度向量对应的覆盖度，选取覆盖度最大的图片作为所述多媒体信息的图片摘要。

9.根据权利要求1所述的图文摘要生成方法，其特征在于，在“基于所述摘要生成模型并根据所述图片对应的覆盖度向量获取所述多媒体信息的图片摘要”的步骤之前，所述方法还包括：

按照下式所示的方法优化所述摘要生成模型的参数：

$$cov_{img}^{t} = \sum_{\tilde{t}=0}^{t-1} \alpha_{img}^{\tilde{t}}$$

$$L_t = -\log\left(p_{w_t}\right) + \sum_{i} \min\left(\alpha_i^t, cov_i^t\right) + \sum_{j} \min\left(\alpha_j^t, cov_{img,j}^t\right)$$

其中，$\alpha_{img}^{\tilde{t}}$表示t时刻的图片特征向量的注意力分布，cov_{img}^{t}表示t时刻的图片覆盖度向量，α_j^t表示t时刻第j个词的图片注意力分布，L_t表示负对数似然损失和覆盖度损失，p_{w_t}表示t时刻从预设的历史词库中生成词或者从输入文本中复制词的概率分布，cov_i^t表示t时刻第i个词的文本覆盖度向量，x_t表示在t时刻的输入的文本词向量。

图文摘要生成方法

技术领域

[0001] 本发明属于自然语言技术领域,具体涉及一种图文摘要生成方法。

背景技术

[0002] 自动摘要是利用计算机系统自动实现文本分析、内容归纳和摘要自动生成的技术,可以按读者(或用户)的要求以简洁的形式表达原文的主要内容。自动摘要技术能够有效地帮助读者(或用户)从检索到的文章中寻找感兴趣的内容,提高阅读速度和质量。该技术可以将文档压缩为更为简洁的表达,并且保证涵盖原始文档有价值的主题。

[0003] 传统的自动摘要技术一般是单模态摘要,即输入全部为文本。随着技术的发展,多模态自动摘要技术出现。多模态自动摘要的输入为多个模态,包括文本、音频、视频和图像等,随着信息的载体越来越丰富多样,当用户通过搜索引擎对某一特定事件进行检索时,返回的内容往往不局限于文本,还可能来源于视频和图像模态。多模态自动摘要技术可以对来自于多模态的信息进行提炼,从而帮助用户在短时间获取多媒体信息。

[0004] 现有的多模态自动摘要技术输出都局限于单模态形式,如只是文本或者图片等,但是实际应用中,文本可以包含准确的语义信息,图片可以帮助用户更快地获取文档主题,这两种模态的信息可以相互补充。现有的方法是将图片和文本作为一个基本的摘要单元联合进行抽取,没有考虑到实际情况中图片和文本都不存在显式的对齐关系,通过这种方式得到的摘要信息是不准确的。

[0005] 因此,如何提出一种将图片与文本对齐从而加速用户获取信息的方案是本领域技术人员目前需要解决的问题。

发明内容

[0006] 为了解决现有技术中的上述问题,即为了解决现有技术图片和文本不对齐导致摘要信息不准确的问题,本发明提供了一种图文摘要生成方法,包括:

[0007] 获取当前获取的多媒体信息中文本和图片对应的特征向量;

[0008] 基于预先构建的多模态信息融合模型并根据所述文本和图片对应的特征向量获取多模态信息向量;

[0009] 基于预先构建的摘要生成模型并根据所述多模态信息向量获取所述多媒体信息的文本摘要;

[0010] 基于预先构建的注意力机制模型并根据图片对应的特征向量获取所述图片对应的覆盖度向量;

[0011] 基于所述摘要生成模型并根据所述图片对应的覆盖度向量获取所述多媒体信息的图片摘要;

[0012] 将所述文本摘要和图片摘要结合作为所述多媒体信息的图文摘要;

[0013] 其中,所述多模态信息融合模型、摘要生成模型以及注意力机制模型均是基于预设的多媒体信息训练数据集并利用机器学习算法所构建的神经网络模型。

[0014] 在上述方案的优选技术方案中，“获取当前获取的多媒体信息中文本和图片对应的特征向量”的步骤包括：

[0015] 根据下式所示的双向长短期记忆网络获取所述多媒体信息中文本的特征向量：

[0016] $f_t=\sigma_g(W_f x_t+U_f c_{t-1}+b_f)$

[0017] $i_t=\sigma_g(W_i x_t+U_i c_{t-1}+b_i)$

[0018] $o_t=\sigma_g(W_o x_t+U_o c_{t-1}+b_o)$

[0019] $c_t=f_t\odot c_{t-1}+i_t\odot\sigma_c(W_c x_t+U_c h_{t-1}+b_c)$

[0020] $h_t=o_t\odot\sigma_h(c_t)$

[0021] 其中，f_t、i_t、o_t分别表示所述双向长短期记忆网络在t时刻的遗忘门、输入门和输出门的输出，σ_g、σ_c、σ_h分别表示遗忘门、输入门和输出门的激活函数，W_f、W_i、W_o分别表示遗忘门、输入门和输出门的第一矩阵参数，U_f、U_i、U_o分别表示遗忘门、输入门和输出门的第二矩阵参数，x_t表示在t时刻的输入的文本词向量，c_{t-1}表示在t-1时刻的文本的特征向量，b_f、b_i、b_o分别表示遗忘门、输入门和输出门的偏置参数，h_t表示文本的特征向量对应的隐层向量；

[0022] 基于预先构建的图片特征提取模型获取所述多媒体信息中图片的fc7特征或者pool5特征，将所述fc7特征或者pool5特征转换为图片对应的特征向量；

[0023] 其中，所述图片特征提取模型是基于预设的图片数据集并利用机器学习算法所构建的神经网络模型。

[0024] 在上述方案的优选技术方案中，“将所述fc7特征或者pool5特征转换为图片对应的特征向量”的步骤包括：

[0025] 将所述fc7特征与预先获取的图片的特征向量的注意力分布相乘，得到所述图片对应的特征向量；或者

[0026] 将所述pool5特征与预先获取的图片的特征向量的注意力分布相乘，得到所述图片对应的特征向量；或者

[0027] 获取图片多个区域的注意力分布，根据所述图片多个区域的注意力分布以及图片多个区域对应的向量进行加权求和，将加权求和的结果与预先获取的图片的特征向量的注意力分布相乘，得到所述图片对应的特征向量。

[0028] 在上述方案的优选技术方案中，“基于预先构建的多模态信息融合模型并根据文本的特征向量和图片的特征向量获取多模态信息向量”的步骤包括：

[0029] 根据下式所述的注意力机制获取所述多模态信息向量：

[0030] $\alpha_{txt}^t=\sigma(W_{txt}c_{txt}^t+U_{txt}s_t)$

[0031] $\alpha_{img}^t=\sigma(W_{img}c_{img}^t+U_{img}s_t)$

[0032] $c_{mm}^t=\alpha_{txt}^t c_{txt}^t+\alpha_{img}^t c_{img}^t$

[0033] 其中，α_{txt}^t、α_{img}^t分别表示文本和图片的特征向量的注意力分布，σ表示激活函数，W_{txt}、W_{img}分别表示所述多模态信息融合模型的第一矩阵参数，c_{txt}^t、c_{img}^t分别表示文本和图片的特征向量，U_{txt}、U_{img}分别表示所述多模态信息融合模型的第二矩阵参数，s_t表示所述多模态信息融合模型的状态参数，c_{mm}^t表示所述多模态信息向量。

[0034] 在上述方案的优选技术方案中，在“基于预先构建的摘要生成模型并根据所述多

模态信息向量获取所述多媒体信息的文本摘要”的步骤之前，所述方法还包括：

[0035] 基于预先获取的多模态信息向量并利用注意力机制计算从预设的历史词库中生成和/或复制所述多模态信息中文本的概率；

[0036] 根据所述概率并利用负对数似然损失函数以及覆盖度损失函数优化所述摘要生成模型的参数。

[0037] 在上述方案的优选技术方案中，“根据所述概率并利用负对数似然损失函数以及覆盖度损失函数优化所述摘要生成模型的参数”的步骤包括：

[0038] 按照下式所示的方法优化所述摘要生成模型的参数：

[0039] $$p_g = \sigma(W_h^* c_{mm} + W_s^* s_t + W_x x_t)$$

[0040] $$p_w = p_g p_v(w) + (1 - p_g) \sum_{w_i = w} \alpha_i^t$$

[0041] $$L_t = -\log(p_{w_t}) + \sum_i \min(\alpha_i^t, cov_i^t)$$

[0042] 其中，p_g表示从预设的历史词库中生成词的概率，σ表示激活函数，W_h^*、W_s^*、W_x均表示摘要生成模型的矩阵参数，c_{mm}表示多模态信息向量，s_t表示摘要生成模型的状态参数，p_w表示一个词生成和/或复制的概率，$p_v(w)$表示从预设的历史词库中生成词w的概率，α_i^t表示t时刻第i个词的文本注意力分布，L_t表示负对数似然损失和覆盖度损失，p_{w_t}表示t时刻从预设的历史词库中生成词或者从输入文本中复制词的概率分布，cov_i^t表示t时刻第i个词的文本覆盖度向量。

[0043] 在上述方案的优选技术方案中，“基于预先构建的注意力机制模型并根据图片对应的特征向量获取所述图片对应的覆盖度向量”的步骤包括：

[0044] 基于所述注意力机制模型获取所述图片对应的特征向量多个时刻的注意力分布，将所述多个时刻的注意力分布累加得到所述图片对应的覆盖度向量。

[0045] 在上述方案的优选技术方案中，“基于所述摘要生成模型并根据所述图片对应的覆盖度向量获取所述多媒体信息的图片摘要”的步骤包括：

[0046] 基于所述摘要生成模型获取每张图片的覆盖度向量对应的覆盖度，选取覆盖度最大的图片作为所述多媒体信息的图片摘要。

[0047] 在上述方案的优选技术方案中，在“基于所述摘要生成模型并根据所述图片对应的覆盖度向量获取所述多媒体信息的图片摘要”的步骤之前，所述方法还包括：

[0048] 按照下式所示的方法优化所述摘要生成模型的参数：

[0049] $$cov_{img}^t = \sum_{\tilde{t}=0}^{t-1} \alpha_{img}^{\tilde{t}}$$

[0050] $$L_t = -\log(p_{w_t}) + \sum_i \min(\alpha_i^t, cov_i^t) + \sum_j \min(\alpha_j^t, cov_{img,j}^t)$$

[0051] 其中，$\alpha_{img}^{\tilde{t}}$表示t时刻的图片特征向量的注意力分布，cov_{img}^t表示t时刻的图片覆盖度向量，α_j^t表示t时刻第j个词的图片注意力分布。

[0052] 与最接近的现有技术相比，上述技术方案至少具有如下有益效果：

[0053] 1、本发明提供的图文摘要生成方法，是利用序列到序列的框架生成文本摘要，结合注意力机制捕捉文本和图片的对齐关系，利用覆盖度机制选出最重要的图片，将文本摘要和图片摘要结合作为最终的图文摘要，通过对齐文本和图片，可以得到更准确表现多媒体信息内容的图文摘要；

[0054] 2、通过预先构建的注意力机制模型并根据图片对应的特征向量获取图片对应的覆盖度向量，根据覆盖度向量获取多媒体信息的图片摘要，可以根据每张图片的覆盖度得到每张图片的重要性分数，将重要性分数最高的图片作为图片摘要，可以使用户能够通过图片更快地获取多媒体信息的主题。

附图说明

[0055] 图1为本发明一种实施例的图文摘要生成方法的主要步骤示意图；

[0056] 图2为本发明实施例中第一种获取图片特征向量的主要步骤示意图；

[0057] 图3为本发明实施例中第二种获取图片特征向量的主要步骤示意图；

[0058] 图4为本发明实施例中第三种获取图片特征向量的主要步骤示意图。

具体实施方式

[0059] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0060] 下面参照附图来描述本发明的优选实施方式。本领域技术人员应当理解的是，这些实施方式仅仅用于解释本发明的技术原理，并非旨在限制本发明的保护范围。

[0061] 参阅附图1，图1示例性的给出了本实施例中图文摘要生成方法的主要步骤。如图1所示，本实施例中图文摘要生成方法包括下述步骤：

[0062] 步骤S101：获取当前获取的多媒体信息中文本和图片对应的特征向量；

[0063] 多媒体信息中文字可以准确地表达语义信息，图片可以帮助用户快速地获取主题，两种模态的信息能够相互补充。为了能够得到对齐的文本和图片，可以获取多媒体信息中文本和图片对应的特征向量。以一条含有M张图片的新闻为例，其中下面两条文本分别是输入文本和人工参考摘要：

[0064] 输入文本：It’s just an example for illustration.

[0065] 人工参考摘要：It’s an example.

[0066] 为了减少后期的计算量，可以将新闻中所有的英文文本以及参考摘要进行分词和小写转换，具体地，可以采用开源的分词工具对英文文档进行分词，以上述给出的内容为例，进行分词和小写转换后，输入文本和人工参考摘要如下所示：

[0067] it’s just an example for illustration.

[0068] it’s an example.

[0069] 对多媒体信息进行预处理后，可以分别获取多媒体信息中文本和图片对应的特征向量，具体地，可以根据公式(1)和公式(2)所示的双向长短期记忆网络获取多媒体信息中

文本的特征向量：

[0070] $$\begin{cases} f_t = \sigma_g(W_f x_t + U_f c_{t-1} + b_f) \\ i_t = \sigma_g(W_i x_t + U_i c_{t-1} + b_i) \\ o_t = \sigma_g(W_o x_t + U_o c_{t-1} + b_o) \end{cases} \quad (1)$$

[0071] $$\begin{cases} c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\ h_t = o_t \odot \sigma_h(c_t) \end{cases} \quad (2)$$

[0072] 其中，f_t、i_t、o_t分别表示双向长短期记忆网络在t时刻的遗忘门、输入门和输出门的输出，σ_g、σ_c、σ_h分别表示遗忘门、输入门和输出门的激活函数，W_f、W_i、W_o分别表示遗忘门、输入门和输出门的第一矩阵参数，U_f、U_i、U_o分别表示遗忘门、输入门和输出门的第二矩阵参数，x_t表示在t时刻的输入的文本词向量，c_{t-1}表示在t-1时刻的文本的特征向量，b_f、b_i、b_o分别表示遗忘门、输入门和输出门的偏置参数，h_t表示文本的特征向量对应的隐层向量；

[0073] 可以基于预先构建的图片特征提取模型获取多媒体信息中图片的fc7特征或者pool5特征，fc7特征和pool5特征分别为4096维向量和49x512维矩阵。将fc7特征或者pool5特征转换为图片对应的特征向量，其中，图片特征提取模型是基于预设的图片数据集并利用机器学习算法所构建的神经网络模型，具体地，图片特征提取模型可以是训练好的VGG19模型，将fc7特征或者pool5特征转换为图片对应的特征向量的步骤可以包括：

[0074] 如图2所示，图2示例性得给出了本实施例中第一种获取图片特征向量的主要步骤，将fc7特征与预先获取的图片的特征向量的注意力分布相乘，得到图片对应的特征向量；或者

[0075] 如图3所示，图3示例性得给出了本实施例中第二种获取图片特征向量的主要步骤，将pool5特征与预先获取的图片的特征向量的注意力分布相乘，得到图片对应的特征向量；或者

[0076] 如图4所示，图4示例性得给出了本实施例中第三种获取图片特征向量的主要步骤，获取图片多个区域的注意力分布，根据图片多个区域的注意力分布以及图片多个区域对应的向量进行加权求和，将加权求和的结果与预先获取的图片的特征向量的注意力分布相乘，得到图片对应的特征向量。

[0077] 步骤S102：基于预先构建的多模态信息融合模型并根据文本和图片对应的特征向量获取多模态信息向量。

[0078] 具体地，可以使用多模态信息融合模型计算输入文本和输入图片的注意力权重，根据注意力权重将文本和图片输入组成为一个多模态信息向量，可以按照公式(3)所示的方法获取多模态信息向量：

[0079] $$\begin{cases} \alpha_{txt}^t = \sigma(W_{txt} c_{txt}^t + U_{txt} s_t) \\ \alpha_{img}^t = \sigma(W_{img} c_{img}^t + U_{img} s_t) \\ c_{mm}^t = \alpha_{txt}^t c_{txt}^t + \alpha_{img}^t c_{img}^t \end{cases} \quad (3)$$

[0080] 其中，α_{txt}^t、α_{img}^t分别表示文本和图片的特征向量的注意力分布，σ表示激活函数，W_{txt}、W_{img}分别表示多模态信息融合模型的第一矩阵参数，c_{txt}^t、c_{img}^t分别表示文本和图片的特征向量，U_{txt}、U_{img}分别表示多模态信息融合模型的第二矩阵参数，s_t表示多模态信息

融合模型的状态参数，c_{mm}^{t}表示多模态信息向量。

[0081] 在实际应用中，为了更好地获得多模态信息向量，可以在获取多模态信息向量前对多模态信息融合模型进行训练，具体地，可以基于预先获取的多模态信息向量并利用注意力机制计算从预设的历史词库中生成和/或复制多模态信息中文本的概率，根据概率并利用负对数似然损失函数以及覆盖度损失函数优化摘要生成模型的参数，具体方法可以按照公式(4)所示的方法训练多模态信息融合模型：

[0082]
$$\begin{cases} p_g = \sigma(W_h^* c_{mm} + W_s^* s_t + W_x x_t) \\ p_w = p_g p_v(w) + (1 - p_g) \sum_{w_i = w} \alpha_i^t \\ L_t = -\log(p_{w_t}) + \sum_i \min(\alpha_i^t, cov_i^t) \end{cases} \quad (4)$$

[0083] 其中，p_g表示从预设的历史词库中生成词的概率，σ表示激活函数，W_h^*、W_s^*、W_x均表示摘要生成模型的矩阵参数，c_{mm}表示多模态信息向量，s_t表示摘要生成模型的状态参数，p_w表示从预设的历史词库中复制词的概率，$p_v(w)$表示从预设的历史词库中复制词w的概率，α_i^t表示t时刻第i个词的文本注意力分布，L_t表示负对数似然损失和覆盖度损失，p_{w_t}表示t时刻从预设的历史词库中复制词的概率分布，cov_i^t表示t时刻第i个词的文本覆盖度向量。

[0084] 步骤S103：基于预先构建的摘要生成模型并根据多模态信息向量获取多媒体信息的文本摘要；

[0085] 在实际应用中，可以根据摘要生成模型和多模态信息向量计算从预设的历史词库中生成和/或复制多模态信息中文本的概率，将多媒体信息中的文本与历史词库中的文本进行比较，判断多媒体信息中的文本是否出现在历史词库中，若出现，则计算从历史词库中生成文本的概率，若未出现，则计算从输入文本中复制该文本的概率，将生成和/或复制文本概率中概率最大的文本作为文本摘要。

[0086] 为了更好地获得文本摘要，在获取多媒体信息的文本摘要之前，可以基于预先获取的多模态信息向量并利用注意力机制计算从预设的历史词库中生成和/或复制多模态信息中文本的概率，根据概率并利用负对数似然损失函数以及覆盖度损失函数优化摘要生成模型的参数，具体方法可以按照公式(5)所示的方法优化摘要生成模型的参数：

[0087]
$$\begin{cases} p_g = \sigma(W_h^* c_{mm} + W_s^* s_t + W_x x_t) \\ p_w = p_g p_v(w) + (1 - p_g) \sum_{w_i = w} \alpha_i^t \\ L_t = -\log(p_{w_t}) + \sum_i \min(\alpha_i^t, cov_i^t) \end{cases} \quad (5)$$

[0088] 其中，p_g表示从预设的历史词库中生成词的概率，σ表示激活函数，W_h^*、W_s^*、W_x均表示摘要生成模型的矩阵参数，c_{mm}表示多模态信息向量，s_t表示摘要生成模型的状态参数，p_w表示一个词生成和/或复制的概率，$p_v(w)$表示从预设的历史词库中生成词w的概率，α_i^t表示t时刻第i个词的文本注意力分布，L_t表示负对数似然损失和覆盖度损失，p_{w_t}表示t时刻从预设的历史词库中生成词或者从输入文本中复制词的概率分布，cov_i^t表示t时刻第i个词的文本覆盖度向量。

[0089] 经过训练后的摘要生成模型可以更准确地获取文本摘要，其中，摘要生成模型可

以是单向循环神经网络。

[0090] 步骤S104:基于预先构建的注意力机制模型并根据图片对应的特征向量获取图片对应的覆盖度向量;

[0091] 在实际应用中,图片可以帮助用户更快地获取文档主题,但是多媒体信息中可能包含多张图片,为了帮助用户尽快地获取文档主题,需要从多媒体信息的多张图片中挑选出最能表现文档主题的图片,具体地,可以通过注意力机制获取每个时刻图片的注意力分布,将多个时刻的图片的注意力分布进行累加,得到图片对应的覆盖度向量,在通过覆盖度损失函数计算得到图片对应的覆盖度向量,其中,图片不同的注意力形式对应于不同的图片重要性的计算方式,可以根据单张图片的覆盖度向量选取覆盖度最大的图片作为摘要图片。

[0092] 为了更好地获得图片摘要,在获取多媒体信息的图片摘要之前,可以进一步优化摘要生成模型的参数,具体方法可以按照公式(6)所示的方法优化摘要生成模型的参数:

[0093]
$$\begin{cases} cov_{img}^{t} = \sum_{\tilde{t}=0}^{t-1} \alpha_{img}^{\tilde{t}} \\ L_t = -\log\left(p_{w_t}\right) + \sum_i \min\left(\alpha_i^t, cov_i^t\right) + \sum_j \min\left(\alpha_j^t, cov_{img,j}^t\right) \end{cases} \quad (6)$$

[0094] 其中,$\alpha_{img}^{\tilde{t}}$表示t时刻的图片特征向量的注意力分布,cov_{img}^{t}表示t时刻的图片覆盖度向量,α_j^t表示t时刻第j个词对应的图片注意力分布。

[0095] 步骤S105:基于摘要生成模型并根据图片对应的覆盖度向量获取多媒体信息的图片摘要。

[0096] 具体地,摘要生成模型可以获取每张图片的覆盖度向量对应的覆盖度,比较每张图片的覆盖度大小,覆盖度越大的,说明其重要性分数越高,越能体现文档的主题,将覆盖度最大的图片作为摘要图片。得到摘要图片后,可以将其与前述步骤得到的文本摘要进行结合,将结合后的图片摘要和文本摘要作为多媒体信息的图文摘要。

[0097] 具体地,附表1给出了本发明与基于序列到序列模型、融合语言特征的序列到序列特征模型以及指针-生成器模型在数据集上单纯考虑文本的ROUGE值。训练数据包含293,965篇新闻文档,其中含有1,928,356张图片;验证集中包含10,355篇新闻文档,其中含有68,520张图片;测试集中包含10,261篇新闻文档,其中含有71,509张图片。本发明实施例给出的参考答案是一段文本摘要加至多三幅相关的图片,都是在测试集上人为标注的。从附表1中可以看出,本发明的多模态的模型在传统的文本摘要的评测中没有明显的优势,而且ROUGE也无法用来评价图文并茂的摘要。

[0098] 附表1:本发明与基于序列到序列模型(S2S+attn),融合语言特征的序列到序列模型(AED)以及指针-生成器模型(PGC)的ROUGE值对比

[0099]

模型	ROUGE-1	ROUGE-2	ROUGE-L
S2S+attn	32.32	12.44	29.65
AED	34.78	13.10	32.24
PGC	**41.11**	**18.31**	**37.74**
ATG(本发明)	40.63	18.12	37.53
ATL(本发明)	**40.86**	18.27	**37.75**
HAN(本发明)	40.82	**18.30**	37.70

[0100] 附表2给出了本发明与指针-生成器模型的人工评价结果，实验结果表明本发明产生的图文摘要能够比较明显的提升用户的满意度。

[0101] 附表2：本发明与指针-生成器模型的人工评价结果

[0102]

模型	PGC	ATG	ATL	HAN
人工评分	3.07	3.30	3.22	3.20

[0103] 上述实施例中虽然将各个步骤按照上述先后次序的方式进行了描述，但是本领域技术人员可以理解，为了实现本实施例的效果，不同的步骤之间不必按照这样的次序执行，其可以同时（并行）执行或以颠倒的次序执行，这些简单的变化都在本发明的保护范围之内。

[0104] 本领域技术人员应该能够意识到，结合本文中所公开的实施例描述的各示例的方法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明电子硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以电子硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。本领域技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0105] 至此，已经结合附图所示的优选实施方式描述了本发明的技术方案，但是，本领域技术人员容易理解的是，本发明的保护范围显然不局限于这些具体实施方式。在不偏离本发明的原理的前提下，本领域技术人员可以对相关技术特征做出等同的更改或替换，这些更改或替换之后的技术方案都将落入本发明的保护范围之内。

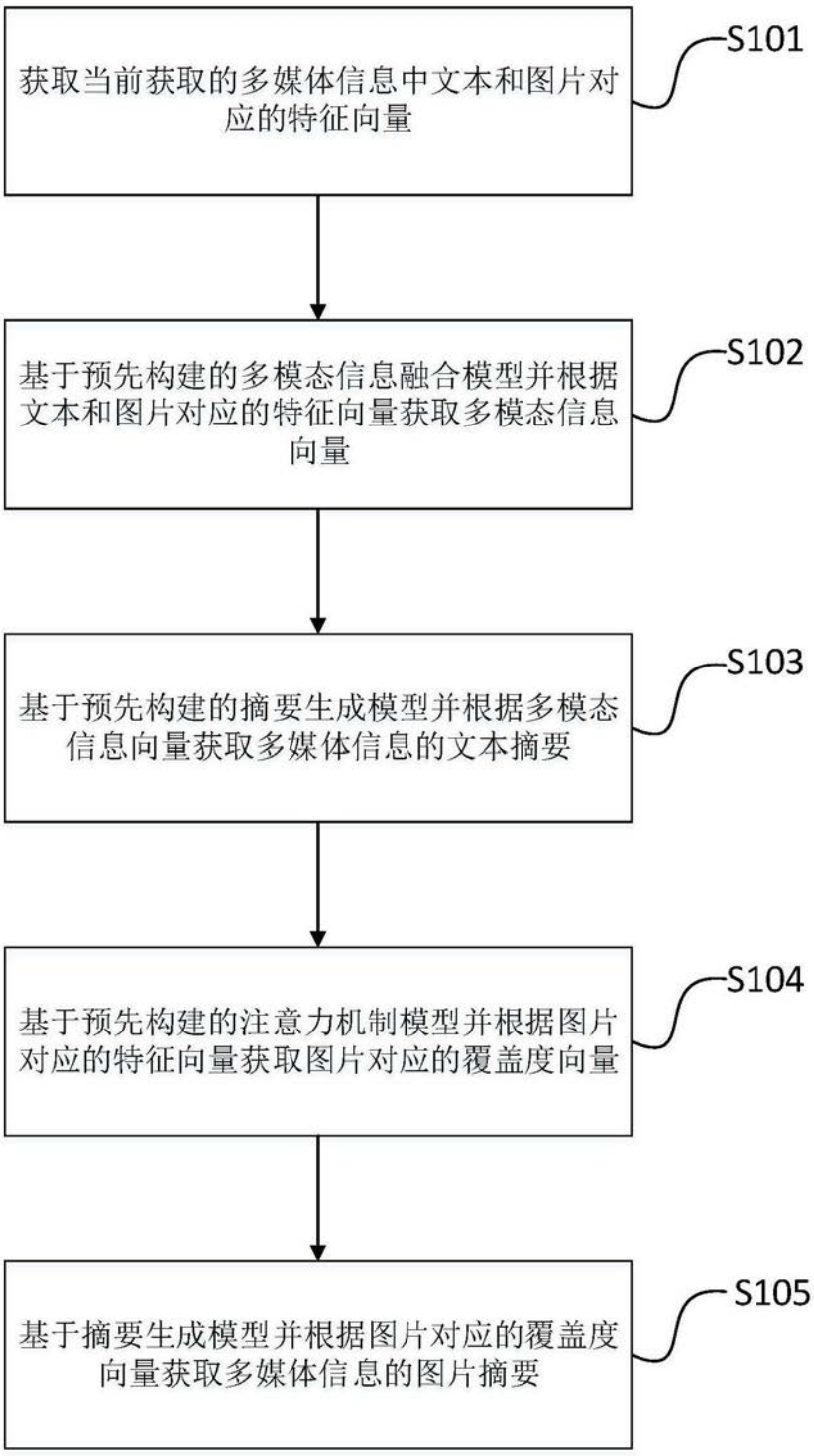

获取当前获取的多媒体信息中文本和图片对应的特征向量
S101
基于预先构建的多模态信息融合模型并根据文本和图片对应的特征向量获取多模态信息向量
S102
基于预先构建的摘要生成模型并根据多模态信息向量获取多媒体信息的文本摘要
S103
基于预先构建的注意力机制模型并根据图片对应的特征向量获取图片对应的覆盖度向量
S104
基于摘要生成模型并根据图片对应的覆盖度向量获取多媒体信息的图片摘要
S105

图1

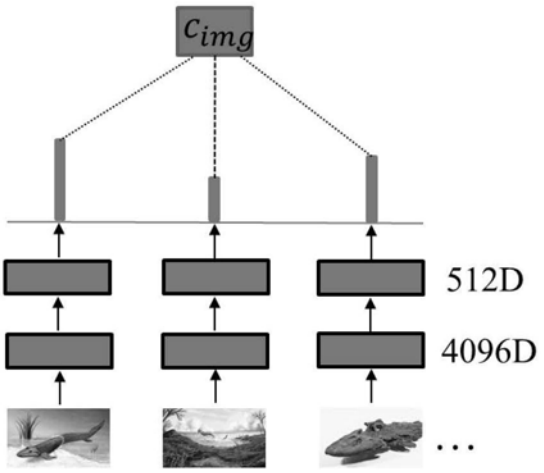


图2

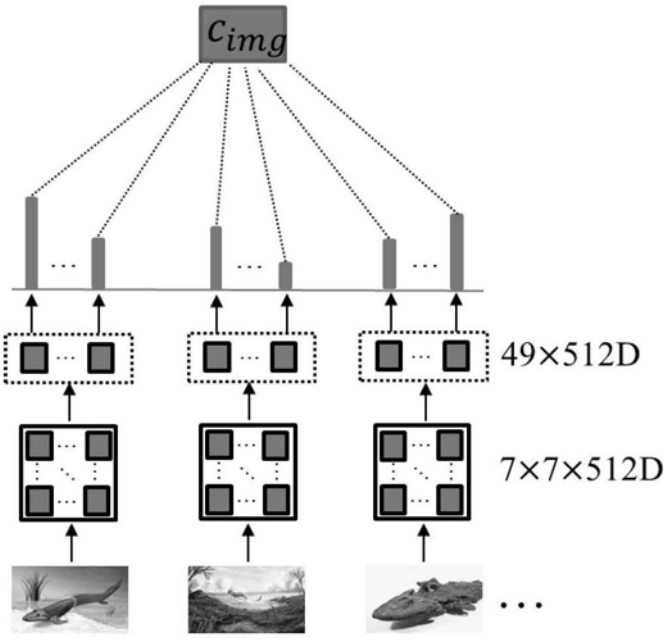


图3

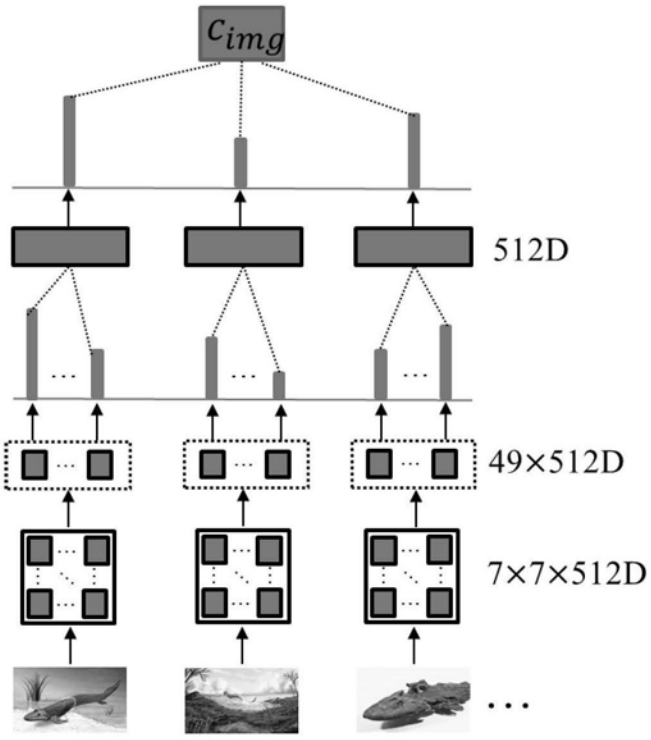


图4