(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114419402 A
(43)申请公布日 2022.04.29

(21)申请号 202210317639.7

(22)申请日 2022.03.29

(71)申请人 中国人民解放军国防科技大学
地址 410073 湖南省长沙市开福区德雅路109号

(72)发明人 谢毓湘 闫洁 宫铨志 魏迎梅 蒋杰 康来 栾悉道 邹诗苇 李竑赋

(74)专利代理机构 长沙国科天河知识产权代理有限公司 43225
代理人 邱轶

(51)Int.Cl.
G06V 10/774(2022.01)
G06F 40/295(2020.01)
G06K 9/62(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

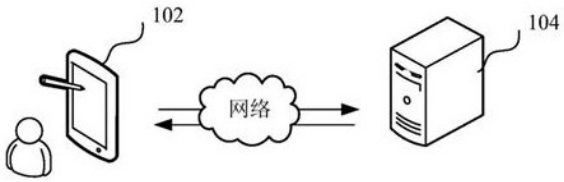
权利要求书2页 说明书10页 附图4页

(54)发明名称

图像故事描述生成方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及一种图像故事描述生成方法、装置、计算机设备和存储介质。所述方法包括：构建数据集；数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个问题描述至少包括疑问词和名词；根据数据集，训练预先构建的图像描述生成模型，以使图像描述生成模型在输入图像时，可以输出图像对应的问题描述；将待描述图像输入训练好的图像描述生成模型，得到待描述图像的问题描述；通过命名实体识别方式从述待描述图像的问题描述中提取疑问词-名词对，将疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本。采用本方法能够更好地指导故事的生成。



CN 114419402 A

1.一种图像故事描述生成方法,其特征在于,所述方法包括:

构建数据集;所述数据集中包括多个图像样本以及每个图像样本对应的问题描述;每个所述问题描述至少包括疑问词和名词;

根据所述数据集,训练预先构建的图像描述生成模型,以使所述图像描述生成模型在输入图像时,可以输出图像对应的问题描述;

将待描述图像输入训练好的图像描述生成模型,得到所述待描述图像的问题描述;

通过命名实体识别方式从所述待描述图像的问题描述中提取疑问词-名词对,将所述疑问词-名词对输入经过预先训练的长文本故事生成模型,得到故事文本;

所述构建数据集,包括:

获取图像样本,确定所述图像样本的疑问词,以及根据所述图像样本,确定与所述图像样本相关联的名词;所述疑问词包括:When、Where、What、Why以及How;

根据每一所述疑问词和对应的所述名词,构建问题描述;所述问题描述包括:When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述;

根据多个图像样本及其对应的所述问题描述,构建数据集。

2.根据权利要求1所述的方法,其特征在于,根据所述数据集,训练预先构建的图像描述生成模型,包括:

将图像样本输入至预先构建的图像描述生成模型中;所述图像描述生成模型包括:特征提取层、编码器和解码器;

通过所述特征提取层对所述图像样本进行特征提取,得到图像特征;

将所述图像特征输入至所述编码器,得到所述图像样本对应的特征向量;

将所述图像样本对应的问题描述进行词嵌入后和所述特征向量分别输入至所述解码器,得到所述特征向量和所述图像样本对应的问题描述进行词嵌入后结果的差值信息;

根据所述差值信息,采用交叉熵损失函数训练预先构建的图像描述生成模型。

3.根据权利要求2所述的方法,其特征在于,所述特征提取层包括:全局特征提取层和局部特征提取层;

所述通过所述特征提取层对所述图像样本进行特征提取,得到图像特征,包括:

通过所述全局特征提取层对所述图像样本进行特征提取,得到全局图像特征;

通过所述局部特征提取层对所述图像样本进行特征提取,得到局部图像特征。

4.根据权利要求3所述的方法,其特征在于,将所述图像特征输入至所述编码器,得到所述图像样本对应的特征向量,包括:

将所述全局图像特征和所述局部图像特征进行拼接融合之后,输出至所述编码器中进行编码,得到所述图像样本对应的特征向量。

5.根据权利要求4中所述的方法,其特征在于,所述全局特征提取层为深度残差网络;所述局部特征提取层为Fast RCNN网络;所述编码器和所述解码器分别为Transformer编码器和Transformer解码器。

6.根据权利要求1至5中任一项所述的方法,其特征在于,训练长文本故事生成模型的方式包括:

通过爬虫从互联网获取英文故事语料库;英文故事语料库包括多个英文故事;

从所述英文故事中提取疑问词-名词对,将英文故事中的疑问词-名词对输入至初始的

长文本故事生成模型中，输出预测故事文本；

根据所述预测故事文本和所述英文故事的差值，采用均方误差损失函数对所述长文本故事生成模型进行训练。

7.根据权利要求2所述的方法，其特征在于，所述根据所述差值信息，采用交叉熵损失函数训练预先构建的图像描述生成模型包括：

根据所述差值信息，得到交叉熵损失函数为：

$$L(\theta)=-\sum_{i=1}^{N}\log\left(p\left(y_i^*\mid y_{1:i-1}^*\right)\right)+\lambda_\theta\|\theta\|_2^2$$

其中，$L(\theta)$表示交叉熵损失函数，θ表示模型中的参数，$p\left(y_i^*\mid y_{1:i-1}^*\right)$表示当前预测输出单词$y_i^*$的概率分布，$y_{1:i-1}^*$表示从第1时刻到第$i-1$时刻所输出的全部单词，$\lambda_\theta\|\theta\|_2^2$表示L2正则化项；

采用所述交叉熵损失函数训练预先构建的图像描述生成模型。

8.一种图像故事描述生成装置，其特征在于，所述装置包括：

数据集构建模块，用于构建数据集；所述数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个所述问题描述至少包括疑问词和名词；

图像描述生成模型训练模块，用于根据所述数据集，训练预先构建的图像描述生成模型，以使所述图像描述生成模型在输入图像时，可以输出图像对应的问题描述；

图像描述生成模块，用于将待描述图像输入训练好的图像描述生成模型，得到所述待描述图像的问题描述；

长文本故事生成模块，用于通过命名实体识别方式从所述待描述图像的问题描述中提取疑问词-名词对，将所述疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本；

数据集构建模块，还用于获取图像样本，确定所述图像样本的疑问词，以及根据所述图像样本，确定与所述图像样本相关联的名词；所述疑问词包括：When、Where、What、Why以及How；根据每一所述疑问词和对应的所述名词，构建问题描述；所述问题描述包括：When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述；根据多个图像样本及其对应的所述问题描述，构建数据集。

9.一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述方法的步骤。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的方法的步骤。

图像故事描述生成方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及多媒体信息处理技术领域，特别是涉及一种图像故事描述生成方法、装置、计算机设备和存储介质。

背景技术

[0002] 随着多媒体信息处理技术的发展，出现了图像描述生成技术，又称为“图像自动注释”，“图像标记”或“图像字幕生成”，是指让计算机根据一幅图像自动生成一段完整而流畅的文字描述声明。图像描述生成任务将计算机视觉和自然语言处理紧密联系在一起，是人工智能领域中的一个基本问题。这项任务会对我们生活的各个方面产生巨大的影响，例如盲人辅助，即帮助视力受损的人更好地理解网络上图像的内容，还可以应用到儿童早教，汽车导航，战场态势分析等实际场景中，以实现更加灵活高效的人机交互。

[0003] 目前关于图像描述的研究主要集中在“生成对图像的白话描述”上，包括提高对图像进行描述的语言的准确性、通俗性、灵活性等。理解一幅图像很大程度上取决于获取图像的特征，用于此目的的技术可大致分为两类：(1)传统的基于机器学习的技术；(2)基于深度学习的技术。传统的基于机器学习的图像描述方法利用了传统的特征提取手段，由于这些手工制作的特征是基于特定任务的，所以用这种方法从大量多样的数据中提取特征是不可行的。此外，真实世界的数据，如图像和视频是复杂的，有不同的语义解释。随着卷积神经网络被广泛用于特征学习，基于深度学习的图像描述生成方法随之流行起来。深度学习是一个端到端的学习过程，可以从训练数据中自动学习特征，因而利用这种方法可以处理大量多样的图像和视频。

[0004] 然而，目前的图像描述故事文本生成方法，存在生成的文本内容不可控且故事性不强的问题。

发明内容

[0005] 基于此，有必要针对上述技术问题，提供一种能够指导长文本故事生成的图像故事描述生成方法、装置、计算机设备和存储介质。

[0006] 一种图像故事描述生成方法，所述方法包括：

构建数据集；所述数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个所述问题描述至少包括疑问词和名词；

根据所述数据集，训练预先构建的图像描述生成模型，以使所述图像描述生成模型在输入图像时，可以输出图像对应的问题描述；

将待描述图像输入训练好的图像描述生成模型，得到所述待描述图像的问题描述；

通过命名实体识别方式从所述待描述图像的问题描述中提取疑问词-名词对，将所述疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本；

所述构建数据集，包括：

获取图像样本，确定所述图像样本的疑问词，以及根据所述图像样本，确定与所述图像样本相关联的名词；所述疑问词包括：When、Where、What、Why以及How；

根据每一所述疑问词和对应的所述名词，构建问题描述；所述问题描述包括：When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述；

根据多个图像样本及其对应的所述问题描述，构建数据集。

[0007] 在其中一个实施例中，还包括：将图像样本输入至预先构建的图像描述生成模型中；所述图像描述生成模型包括特征提取层、编码器和解码器；通过所述特征提取层对所述图像样本进行特征提取，得到图像特征；将所述图像特征输入至所述编码器，得到所述图像样本对应的特征向量；将所述图像样本对应的问题描述进行词嵌入后和所述特征向量分别输入至所述解码器，得到所述特征向量和所述图像样本对应的问题描述进行词嵌入后结果的差值信息；根据所述差值信息，采用交叉熵损失函数训练预先构建的图像描述生成模型。

[0008] 在其中一个实施例中，还包括：特征提取层包括全局特征提取层和局部特征提取层；通过所述特征提取层对所述图像样本进行特征提取，得到图像特征，通过所述全局特征提取层对所述图像样本进行特征提取，得到全局图像特征；通过所述局部特征提取层对所述图像样本进行特征提取，得到局部图像特征。

[0009] 在其中一个实施例中，还包括：将所述全局图像特征和所述局部图像特征进行拼接融合之后，输出至所述编码器中进行编码，得到所述图像样本对应的特征向量。

[0010] 在其中一个实施例中，还包括：全局特征提取层为深度残差网络；所述局部特征提取层为Fast RCNN网络，所述编码器和所述解码器分别为Transformer编码器和Transformer解码器。

[0011] 在其中一个实施例中，还包括：通过爬虫从互联网获取英文故事语料库；英文故事语料库包括多个英文故事；从所述英文故事中提取疑问词-名词对，将英文故事中的疑问词-名词对输入至初始的长文本故事生成模型中，输出预测故事文本；根据所述预测故事文本和所述英文故事的差值，采用均方误差损失函数对所述长文本故事生成模型进行训练。

[0012] 在其中一个实施例中，还包括：根据所述差值信息，得到交叉熵损失函数为：

$$L(\theta)=-\sum_{i=1}^{N}\log\left(p\left(y_i^*\mid y_{1:i-1}^*\right)\right)+\lambda_\theta\|\theta\|_2^2$$

其中，$L(\theta)$表示交叉熵损失函数，θ表示模型中的参数，$p\left(y_i^*\mid y_{1:i-1}^*\right)$表示当前预测输出单词$y_i^*$的概率分布，$y_{1:i-1}^*$表示从第1时刻到第$i-1$时刻所输出的全部单词，$\lambda_\theta\|\theta\|_2^2$表示L2正则化项；采用所述交叉熵损失函数训练预先构建的图像描述生成模型。

[0013] 一种图像故事描述生成装置，所述装置包括：

数据集构建模块，用于构建数据集；所述数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个所述问题描述至少包括疑问词和名词；

图像描述生成模型训练模块，用于根据所述数据集，训练预先构建的图像描述生成模型，以使所述图像描述生成模型在输入图像时，可以输出图像对应的问题描述；

图像描述生成模块，用于将待描述图像输入训练好的图像描述生成模型，得到所述待描述图像的问题描述；

长文本故事生成模块，用于通过命名实体识别方式从所述待描述图像的问题描述

中提取疑问词-名词对，将所述疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本；

数据集构建模块，还用于获取图像样本，确定所述图像样本的疑问词，以及根据所述图像样本，确定与所述图像样本相关联的名词；所述疑问词包括：When、Where、What、Why以及How；根据每一所述疑问词和对应的所述名词，构建问题描述；所述问题描述包括：When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述；根据多个图像样本及其对应的所述问题描述，构建数据集。

[0014] 一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现以下步骤：

构建数据集；所述数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个所述问题描述至少包括疑问词和名词；

根据所述数据集，训练预先构建的图像描述生成模型，以使所述图像描述生成模型在输入图像时，可以输出图像对应的问题描述；

将待描述图像输入训练好的图像描述生成模型，得到所述待描述图像的问题描述；

通过命名实体识别方式从所述待描述图像的问题描述中提取疑问词-名词对，将所述疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本；

所述构建数据集，包括：

获取图像样本，确定所述图像样本的疑问词，以及根据所述图像样本，确定与所述图像样本相关联的名词；所述疑问词包括：When、Where、What、Why以及How；

根据每一所述疑问词和对应的所述名词，构建问题描述；所述问题描述包括：When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述；

根据多个图像样本及其对应的所述问题描述，构建数据集。

[0015] 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现以下步骤：

构建数据集；所述数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个所述问题描述至少包括疑问词和名词；

根据所述数据集，训练预先构建的图像描述生成模型，以使所述图像描述生成模型在输入图像时，可以输出图像对应的问题描述；

将待描述图像输入训练好的图像描述生成模型，得到所述待描述图像的问题描述；

通过命名实体识别方式从所述待描述图像的问题描述中提取疑问词-名词对，将所述疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本；

所述构建数据集，包括：

获取图像样本，确定所述图像样本的疑问词，以及根据所述图像样本，确定与所述图像样本相关联的名词；所述疑问词包括：When、Where、What、Why以及How；

根据每一所述疑问词和对应的所述名词，构建问题描述；所述问题描述包括：When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述；

根据多个图像样本及其对应的所述问题描述，构建数据集。

[0016] 上述图像故事描述生成方法、装置、计算机设备和存储介质，通过获取待描述图像，将待描述图像输入预先训练好的图像描述生成模型，可以得到待描述图像的问题描述，从而使生成的故事文本具有逻辑性，通过命名实体识别方式识别生成的问题描述，能够提取出疑问词-名词对，并将疑问词-名词对输入预先训练好的长文本故事生成模型，得到与待描述图像对应的故事文本，其中，图像描述生成模型基于多个图像样本以及每个图像样本对应的问题描述构建的数据集训练得到，长文本故事生成模型基于爬虫从互联网获取英文故事语料库训练得到，基于所述图像故事描述生成方法，能够更好地指导长文本故事的生成。

附图说明

[0017] 图1为一个实施例中图像故事描述生成方法的应用场景图；

图2为一个实施例中图像故事描述生成方法的流程示意图；

图3为一个实施例中图像描述生成模型的训练集示意图；

图4为一个具体实施例中图像故事描述生成方法的总体框架图；

图5为一个实施例中图像描述生成模型的模型示意图；

图6为一个实施例中长文本生成模型的模型示意图；

图7为一个实施例中图像故事描述生成设备的结构框图；

图8为一个实施例中计算机设备的内部结构图。

具体实施方式

[0018] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0019] 本申请提供的图像故事描述生成方法，可以应用于如图1所示的应用环境中。其中，终端102通过网络与服务器104进行通信。服务器响应终端的图像故事描述生成请求，根据图像故事描述生成请求，获取待描述图像，将待描述图像输入预先训练好的图像描述生成模型，得到待描述图像的问题描述，通过命名实体识别方式识别生成的问题描述，提取出疑问词-名词对，并将疑问词-名词对输入预先训练好的长文本故事生成模型，得到与待描述图像对应的故事文本，其中，图像描述生成模型基于多个图像样本以及每个图像样本对应的问题描述构建的数据集训练得到，长文本故事生成模型基于爬虫从互联网获取英文故事语料库训练得到，将生成的故事文本反馈至终端102。其中，终端102可以但不限于是各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器104可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

[0020] 在一个实施例中，如图2所示，提供了一种图像故事描述生成方法，以该方法应用于图1中的服务器为例进行说明，包括以下步骤：

步骤202，构建数据集。

[0021] 数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个问题描述至少包括疑问词和名词，问题描述是指描述图像的问句，包括疑问词和名词，问题描述以Caption[n](n=1,2,3,…,n∈N)定义，其中，疑问词可以是When，Where，What，Why，How，名词

可以是图像样本上存在的要素，也可以是通过联想学习方式得到的与图像上的要素相关的名词。以一个图像样本与其对应的问题描述为例对问题描述进行具体说明，以图3为例，图3对应的一组问题描述如下：

Caption[1]:When is the picture taken

Caption[2]:Where is the ocean

Caption[3]:What's in the ship

Caption[4]:Why is the ship in this sea area

Caption[5]:How many people are on board

需要注意的是，数据集中并不给出每个问题描述的具体答案，这些问题描述是为了训练图像描述生成模型，使得当计算机处理从来没有见到过的图像时，能够通过训练好的图像描述生成模型生成类似的问题描述。

[0022] 步骤204，根据数据集，训练预先构建的图像描述生成模型，以使图像描述生成模型在输入图像时，可以输出图像对应的问题描述。

[0023] 训练图像描述生成模型是为了得到一个通用的图像描述生成模型，使得当输入一张数据集中不包含的新图像到计算机中时，利用该模型能够自动生成与所输入图像相关的问题描述。预先构建的图像描述生成模型基于Transformer模型建立。Transformer包括编码组件和解码组件。

[0024] 步骤206，将待描述图像输入图像描述生成模型，得到待描述图像的问题描述。

[0025] 步骤208，通过命名实体识别方式从所述待描述图像的问题描述中提取疑问词-名词对，将疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本。

[0026] 命名实体识别方式(Named Entity Recognition,NER)就是从非结构化的输入文本中抽取出上述实体，并且可以按照业务需求识别出更多类别的实体，命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等，实体这个概念可以很广，只要是业务需要的特殊文本片段都可以称为实体，比如产品名称、型号、价格等。

[0027] 将步骤206得到的每个问题描述作为NER的输入文本，NER抽取出的实体为每个问题描述对应的疑问词-名词对，将疑问词-名词对定义为Pair [n](n=1,2,3,…,n∈N)，以图3为例，从中抽取出来的疑问词-名词对分别为：

Pair [1]: (when, picture)

Pair [2]: (where, ocean)

Pair [3]: (what, ship)

Pair [4]: (why, sea)

Pair [5]: (how, people)

每个问题描述中的疑问词与名词以随机的方式进行组合，当问题描述中有多个名词时，仅抽取问题描述中的主语。疑问词-名词对用以输入预先训练好的长文本故事生成模型指导故事的生成。长文本故事生成模型是基于GPT-2(Generative Pre-Training，生成式的预训练)语言模型构建的，GPT-2语言模型是通用的NLP(Natural Language Processing，自然语言处理)模型，可以生成连贯的文本段落，并且能在未经预训练的情况下，完成阅读理解、问答、机器翻译等多项不同的语言建模任务。训练好的长文本故事生成模型是由从互

联网上爬取的英文故事库微调预先训练的GPT-2模型得到的，预先训练的长文本生成模型能够将输入的疑问词-名词对经过语言建模生成一则与待描述图像相关的故事文本。以图3的问题描述为例，根据从中抽取出来的疑问词-名词对生成的长文本故事示例如下：

We have no idea when the picture was taken, but the ship in it was obviously sailing in a part of the Atlantic Ocean. The sky was gray, and the sea was surging and slapping the ship. The reason why the people on board walked anxiously was that they were confused about where the journey would end. It seemed that everyone's fate depends on this endless sea area. Therefore, they didn't know what to do and how to calm themselves down.

步骤202还包括获取图像样本，确定图像样本的疑问词，以及根据图像样本，确定与图像样本相关联的名词；疑问词包括：When、Where、What、Why以及How；根据每一疑问词和对应的名词，构建问题描述；问题描述包括：When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述；根据多个图像样本及其对应的所述问题描述，构建数据集。

[0028] 数据集的构建基于联想学习方式，联想学习是学习的一种形式。其基本假设是，两个事件A和B在一起的经验使人在它们的内部特征之间建立联想，这种联想会由于各种原因在强度上发生变化，从而影响当A进入意识时，回忆起B的可能性与速度。本发明在实现指导长文本故事生成的过程中，联想学习主要表现在：当人们看到一幅图像的时候，由于长期受到周围社会和自然环境，以及生活经验的影响，能够基于所看到的图像进行一系列的联想。具体应用于构建数据集时对每一图像样本进行联想并通过问题描述的方式表达。从而使生成的故事文本具有逻辑性，且通过词语之间的强关联性和语句之间强关联性使得生成的故事文本具备良好的故事性，更易容易引起用户的共鸣，使得用户能够代入图像所描述的场景中。数据集用于训练图像描述生成模型。

[0029] 上述图像故事描述生成方法，通过获取待描述图像，将待描述图像输入预先训练好的图像描述生成模型，可以得到待描述图像的问题描述，通过命名实体识别方式识别生成的问题描述，能够提取出疑问词-名词对，并将疑问词-名词对输入预先训练好的长文本故事生成模型，得到与待描述图像对应的故事文本，其中，图像描述生成模型基于多个图像样本以及每个图像样本对应的问题描述构建的数据集训练得到，长文本故事生成模型基于爬虫从互联网获取英文故事语料库训练得到，基于图像故事描述生成方法，能够更好地指导长文本故事的生成。

[0030] 在其中一个实施例中，如图5所示，提供一种图像描述生成模型的模型示意图，根据数据集，训练预先构建的图像描述生成模型，包括：将图像样本输入至预先构建的图像描述生成模型中，图像描述生成模型包括特征提取层、编码器和解码器，通过特征提取层对图像样本进行特征提取，得到图像特征，将图像特征输入至编码器，得到图像样本对应的特征向量，将图像样本对应的问题描述进行词嵌入后和特征向量分别输入至解码器，得到特征向量和图像样本对应的问题描述进行词嵌入后结果的差值信息，根据差值信息，采用交叉熵损失函数训练预先构建的图像描述生成模型。

[0031] 在本实施例中，解码器的输入有两部分，第一部分输入为将获取到的图像样本的图像特征输入编码器后得到的特征向量；第二部分输入为将问题描述进行词嵌入后得到的词向量。利用输入解码器后得到的差值信息定义交叉熵损失函数，从而推理图像描述生成

模型，经过数据集训练后使得训练好的图像描述生成模型能够对输入的陌生图像，生成对应的问题描述。

[0032] 需要注意的是，图5所代表的流程图，是技术方案以图3为例形成的实施例，问题描述随输入的图像样本的变化而改变。

[0033] 在其中一个实施例中，特征提取层包括全局特征提取层和局部特征提取层，通过全局特征提取层对图像样本进行特征提取，得到全局图像特征，通过局部特征提取层对图像样本进行特征提取，得到局部图像特征。在本实施例中，全局特征是指图像的整体属性，包括颜色特征、纹理特征和形状特征，比如强度直方图等。由于是像素级的低层可视特征，因此，全局特征具有良好的不变性、计算简单、表示直观等特点，此外，全局特征描述不适用于图像混叠和有遮挡的情况；局部特征则是从图像局部区域中抽取的特征，包括边缘、角点、线、曲线和特别属性的区域等。常见的局部特征包括角点类和区域类两大类描述方式。与线特征、纹理特征、结构特征等全局图像特征相比，局部图像特征具有在图像中蕴含数量丰富，特征间相关度小，遮挡情况下不会因为部分特征的消失而影响其他特征的检测和匹配等特点。提取全局特征与局部特征是为了得到融合特征，将两个特征融合后得到的融合特征能够得到更多的图像信息。

[0034] 在其中一个实施例中，将图像特征输入至编码器，得到图像样本对应的特征向量，包括：将全局图像特征和局部图像特征进行拼接融合之后，输出至编码器中进行编码，得到图像样本对应的特征向量。在本实施例中，通过向量拼接的方式对全局图像特征向量与局部图像特征向量进行特征融合，特征融合的目的，是把从图像中提取的特征，合并成一个比输入特征更具有判别能力的特征。融合特征是图像更丰富的、更加细粒度的特征。

[0035] 在其中一个实施例中，全局特征提取层为深度残差网络，局部特征提取层为Fast RCNN网络，编码器和解码器分别为Transformer编码器和Transformer解码器。在本实施例中，深度残差网络(Deep residual network, ResNet)通过残差学习解决了深度网络的退化问题，可以训练出更深的网络，应用于提取图像样本的全局特征；Fast RCNN网络(Fast Region-based Convolutional Network，快速的基于区域的卷积神经网络)是一种快速的基于区域的卷积网络方法，用于目标检测，应用于提取图像样本的局部特征。

[0036] 在另一个实施例中，如图6所示，提供一种长文本生成模型的模型示意图，训练长文本故事生成模型的方式为：通过爬虫从互联网获取英文故事语料库；英文故事语料库包括多个英文故事，从英文故事中提取疑问词-名词对，将英文故事中的疑问词-名词对输入至初始的长文本故事生成模型中，输出预测故事文本，根据预测故事文本和英文故事的差值，采用均方误差损失函数对长文本故事生成模型进行训练。

[0037] 具体地，用于训练长文本故事生成模型的英文故事语料库的大小大于20MB，长文本故事模型输出的故事样本为文本长度不小于50个单词的英文故事。

[0038] 在其中一个实施例中，根据差值信息，采用交叉熵损失函数训练预先构建的图像描述生成模型包括：根据差值信息，得到交叉熵损失函数为：

$$L(\theta)=-\sum_{i=1}^{N}\log\left(p\left(y_i^*\mid y_{1:i-1}^*\right)\right)+\lambda_\theta\|\theta\|_2^2$$

其中，$L(\theta)$表示交叉熵损失函数，θ表示模型中的参数，$p\left(y_i^*\mid y_{1:i-1}^*\right)$表示当前预测

输出单词y_i^*的概率分布，$y_{1:i-1}^*$表示从第1时刻到第i-1时刻所输出的全部单词，$\lambda_\theta\|\theta\|_2^2$表示L2正则化项；采用交叉熵损失函数训练预先构建的图像描述生成模型。

[0039] 在一个具体实施例中，如图4所示，提供一种图像故事描述生成方法的总体框架图，将待描述图像输入至训练好的图像描述生成模型中，得到图像对应的问题描述分别为Caption[1]、Caption[2]、Caption[3]、 Caption[4]、Caption[5]，图中用问题描述表示Caption，通过命名体识别方式抽取问题描述中的疑问词-名词对为Pair [1]、Pair [2]、Pair [3]、Pair [4]和Pair[5]，图中用疑问词-名词对表示Pair，将抽取出来的疑问词-名词对输入训练好的长文本故事生成模型，得到长文本故事。

[0040] 应该理解的是，虽然图1-6的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图1-6中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0041] 在一个实施例中，如图7所示，提供了一种图像故事描述生成装置，包括：数据集构建模块702、图像描述生成模型训练模块704、图像描述生成模块706和长文本故事生成模块708，其中：

数据集构建模块702，用于构建数据集；数据集中包括多个图像样本以及每个图像样本对应的问题描述；每个问题描述至少包括疑问词和名词；

图像描述生成模型训练模块704，用于根据数据集，训练预先构建的图像描述生成模型，以使图像描述生成模型在输入图像时，可以输出图像对应的问题描述；

图像描述生成模块706，用于将待描述图像输入图像描述生成模型，得到待描述图像的问题描述；

长文本故事生成模块708，用于通过命名实体识别方式从所述待描述图像的问题描述中提取疑问词-名词对，将疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本。

[0042] 数据集构建模块702还用于获取图像样本，确定图像样本的疑问词，以及根据图像样本，确定与图像样本相关联的名词；疑问词包括：When、Where、What、Why以及How，根据每一疑问词和对应的名词，构建问题描述，问题描述包括：When问题描述、Where问题描述、What问题描述、Why问题描述以及How问题描述，根据多个图像样本及其对应的问题描述，构建数据集。

[0043] 在其中一个实施例中，图像描述生成模型训练模块704还用于将图像样本输入至预先构建的图像描述生成模型中，图像描述生成模型包括：特征提取层、编码器和解码器，通过特征提取层对图像样本进行特征提取，得到图像特征，将图像特征输入至编码器，得到图像样本对应的特征向量，将图像样本对应的问题描述进行词嵌入后和特征向量分别输入至解码器，得到特征向量和图像样本对应的问题描述进行词嵌入后结果的差值信息，根据差值信息，采用交叉熵损失函数训练预先构建的图像描述生成模型。

[0044] 在其中一个实施例中，图像描述生成模型训练模块704还用于通过特征提取层对

图像样本进行特征提取，得到图像特征，通过全局特征提取层对图像样本进行特征提取，得到全局图像特征，通过局部特征提取层对图像样本进行特征提取，得到局部图像特征。

[0045] 在其中一个实施例中，图像描述生成模型训练模块704还用于将图像特征输入至编码器，得到图像样本对应的特征向量，将全局图像特征和局部图像特征进行拼接融合之后，输出至编码器中进行编码，得到图像样本对应的特征向量。

[0046] 在其中一个实施例中，图像描述生成模型训练模块704还用于全局特征提取层为深度残差网络，局部特征提取层为Fast RCNN网络，编码器和解码器分别为Transformer编码器和Transformer解码器。

[0047] 在其中一个实施例中，长文本故事生成模块708还用于训练长文本故事生成模型的方式包括：通过爬虫从互联网获取英文故事语料库，英文故事语料库包括多个英文故事，从所述英文故事中提取疑问词-名词对，将英文故事中的疑问词-名词对输入至初始的长文本故事生成模型中，输出预测故事文本，根据所述预测故事文本和所述英文故事的差值，采用均方误差损失函数对所述长文本故事生成模型进行训练。

[0048] 在其中一个实施例中，图像描述生成模型训练模块704还用于根据差值信息，得到交叉熵损失函数为：

$$L(\theta)=-\sum_{i=1}^{N}\log\left(p\left(y_i^*\mid y_{1:i-1}^*\right)\right)+\lambda_\theta\|\theta\|_2^2$$

其中，$L(\theta)$表示交叉熵损失函数，θ表示模型中的参数，$p\left(y_i^*\mid y_{1:i-1}^*\right)$表示当前预测输出单词$y_i^*$的概率分布，$y_{1:i-1}^*$表示从第1时刻到第i-1时刻所输出的全部单词，$\lambda_\theta\|\theta\|_2^2$表示L2正则化项；采用交叉熵损失函数训练预先构建的图像描述生成模型。

[0049] 关于图像故事描述生成装置的具体限定可以参见上文中对于图像故事描述生成方法的限定，在此不再赘述。上述图像故事描述生成装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

[0050] 在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图8所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储图像故事描述生成数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种图像故事描述生成方法。

[0051] 本领域技术人员可以理解，图8中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

[0052] 在一个实施例中，提供了一种计算机设备，包括存储器和处理器，该存储器存储有计算机程序，该处理器执行计算机程序时实现上述实施例中方法的步骤。

[0053] 在一个实施例中，提供了一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现上述实施例中方法的步骤。

[0054] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink) DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0055] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0056] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

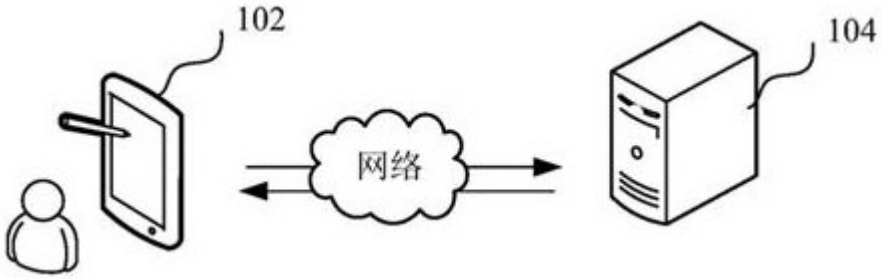

102
104
网络

图1

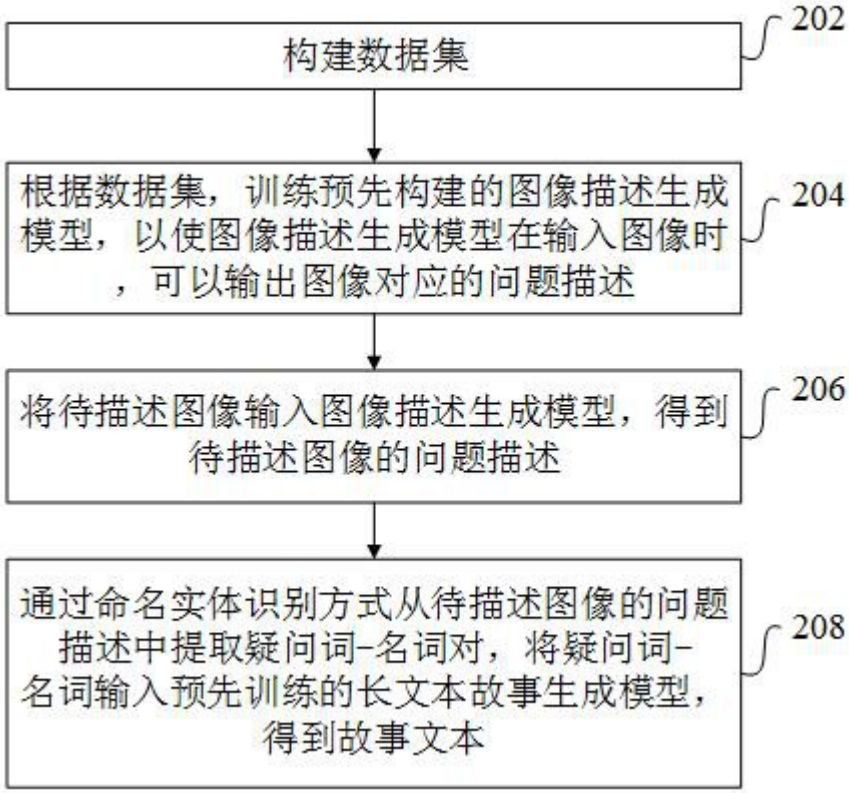

构建数据集
202
根据数据集，训练预先构建的图像描述生成模型，以使图像描述生成模型在输入图像时，可以输出图像对应的问题描述
204
将待描述图像输入图像描述生成模型，得到待描述图像的问题描述
206
通过命名实体识别方式从待描述图像的问题描述中提取疑问词-名词对，将疑问词-名词输入预先训练的长文本故事生成模型，得到故事文本
208

图2

图3

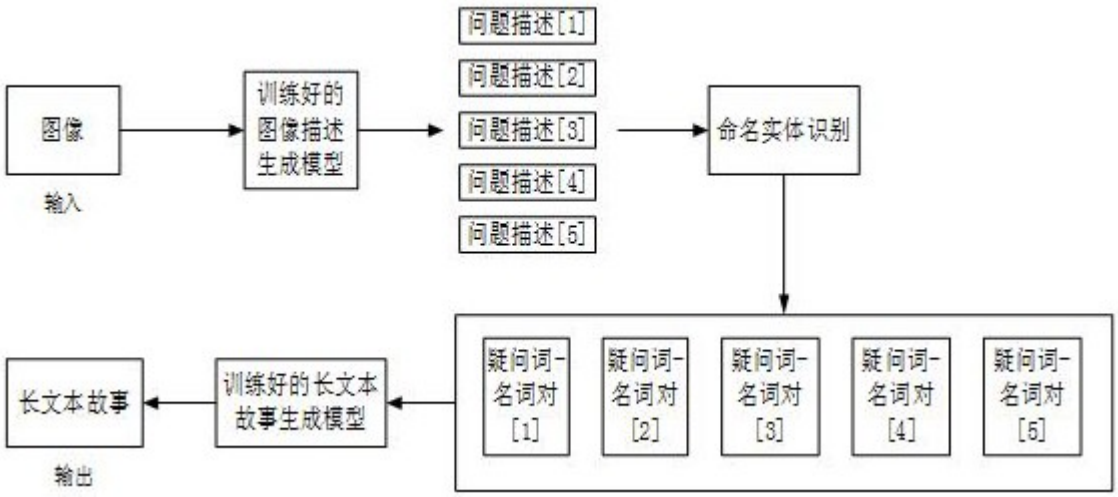

图像
输入
训练好的图像描述生成模型
问题描述[1]
问题描述[2]
问题描述[3]
问题描述[4]
问题描述[5]
命名实体识别
疑问词-名词对[1]
疑问词-名词对[2]
疑问词-名词对[3]
疑问词-名词对[4]
疑问词-名词对[5]
训练好的长文本故事生成模型
长文本故事
输出

图4

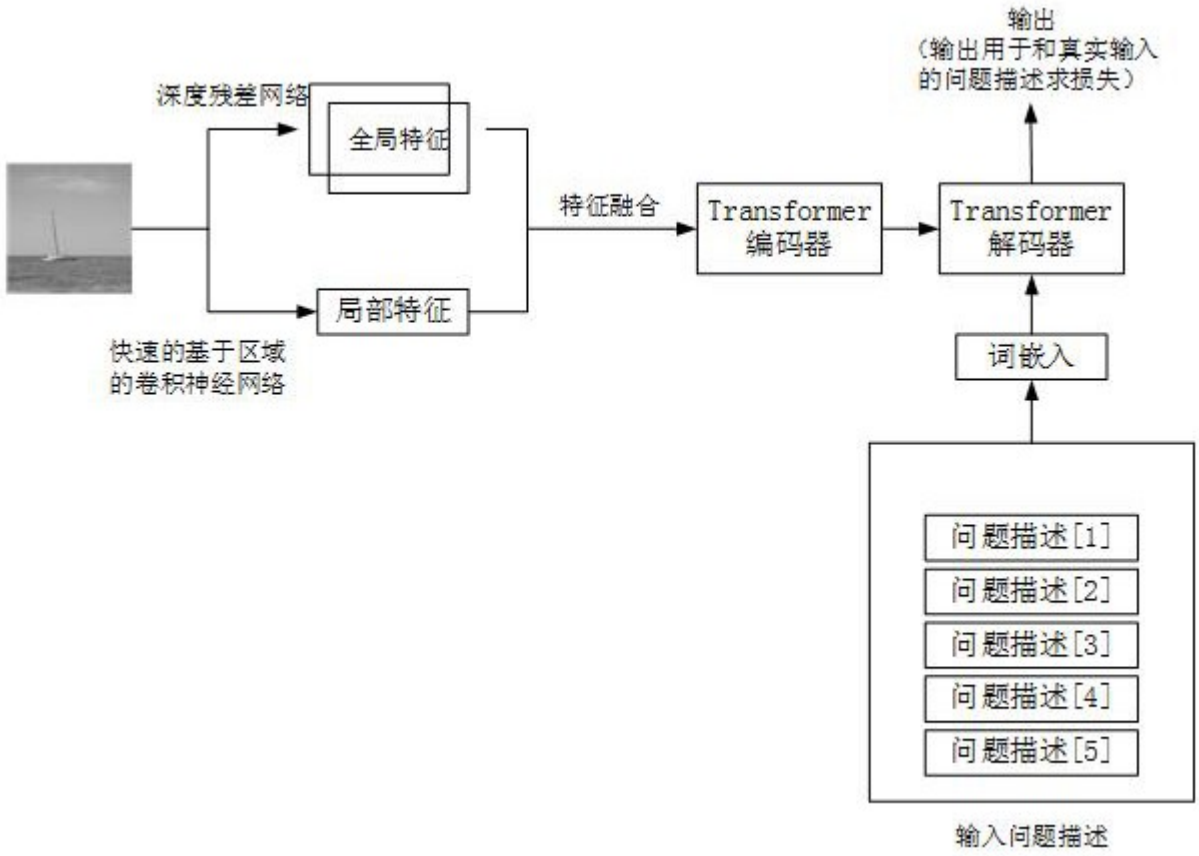

输出
（输出用于和真实输入的问题描述求损失）
深度残差网络
全局特征
局部特征
快速的基于区域的卷积神经网络
特征融合
Transformer编码器
Transformer解码器
词嵌入
问题描述[1]
问题描述[2]
问题描述[3]
问题描述[4]
问题描述[5]
输入问题描述

图5

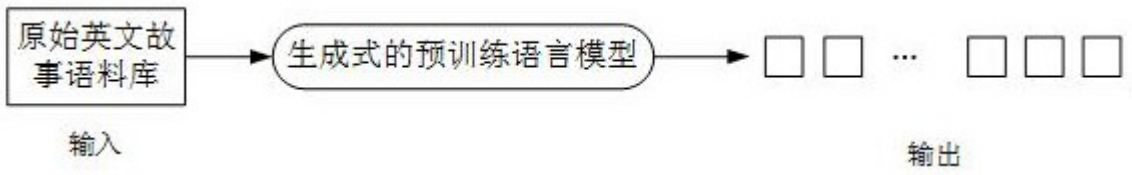

原始英文故事语料库
生成式的预训练语言模型
输入
输出

图6

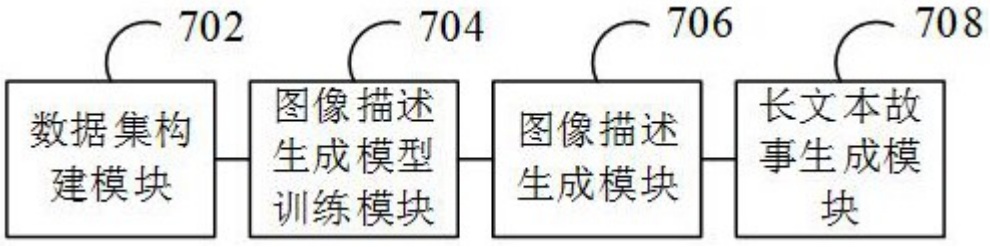

702
704
706
708
数据集构建模块
图像描述生成模型训练模块
图像描述生成模块
长文本故事生成模块

图7

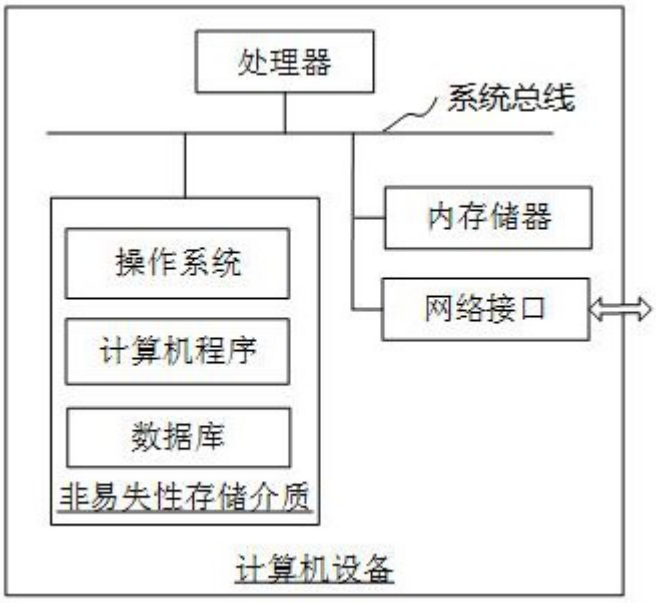


图8