(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 110516791 A
(43)申请公布日 2019.11.29

(21)申请号 201910770172.X

(22)申请日 2019.08.20

(71)申请人 北京影谱科技股份有限公司
地址 100000 北京市朝阳区朝外大街22号5层521室

(72)发明人 刘伟

(74)专利代理机构 北京万思博知识产权代理有限公司 11694
代理人 孙黎生

(51)Int.Cl.
G06N 3/04(2006.01)
G06N 5/04(2006.01)

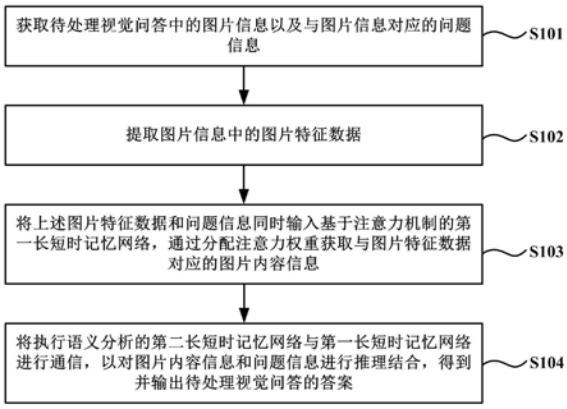
权利要求书2页 说明书9页 附图4页

(54)发明名称

一种基于多重注意力的视觉问答方法及系统

(57)摘要

本申请公开了一种基于多重注意力的视觉问答方法及系统，在本申请提供的方法中，先获取待处理视觉问答中的图片信息及对应的问题信息，提取图片信息中的图片特征数据，然后将图片特征数据和问题信息同时输入基于注意力机制的第一长短时记忆网络，通过分配注意力权重获取图片内容信息，再通过两个双向长短时记忆网络完成问题和图片的结合，输出待处理视觉问答的答案。基于本申请提供的基于多重注意力的视觉问答方法及系统，在R-CNN网络的计算过程中增加一些记忆模块以提高模型在训练过程中的知识源，产生更多样化且合理的答案，端到端的记忆网络实现提取问题信息的同时融合与问题相关的先验知识，提高问题回答上的整体准确率。



CN 110516791 A

1.一种基于多重注意力的视觉问答方法,包括:

获取待处理视觉问答中的图片信息以及与所述图片信息对应的问题信息;

提取所述图片信息中的图片特征数据;

将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络,通过分配注意力权重获取与所述图片特征数据对应的图片内容信息;

将执行语义分析的第二长短时记忆网络与所述第一长短时记忆网络进行通信,以对所述图片内容信息和所述问题信息进行推理结合,得到并输出所述待处理视觉问答的答案。

2.根据权利要求1所述的方法,其特征在于,所述提取所述图片信息中的图片特征数据,包括:

使用区域卷积神经网络R-CNN提取所述图片信息中的至少一个特征区域;

通过ResNet提取每个所述特征区域中的至少一个区域特征信息;

对于每个特征区域,根据重叠度IoU对该特征区域中的区域特征信息进行特征筛选,并将筛选后的区域特征信息作平均池化,进而得到每个特征区域的区域特征数据。

3.根据权利要求2所述的方法,其特征在于,所述提取所述图片信息中的图片特征数据之前,还包括:

基于预设数据集对所述R-CNN和/或ResNet进行预训练;

将所述预设数据集中所包含图片的每个区域特征与代表真实类别的向量融合;

将融合后的向量传输至所述R-CNN和/或ResNet的全连接层输出做属性类别和非属性类别的softmax分类。

4.根据权利要求2所述的方法,其特征在于,所述将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络,通过分配注意力权重获取与所述图片特征数据对应的图片内容信息,包括:

将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络;其中,在所述第一长短时记忆网络的每个时间戳上,其输入包括:上一个时间戳的第二长短时记忆网络的输出、各区域特征数据和上一个时间戳第一长短时记忆网络的输出;其输出包括:对每个所述区域特征数据分配注意力权重;

基于每个所述区域特征数据注意力权重获取与每个所述区域特征数据对应的内容信息。

5.根据权利要求4所述的方法,其特征在于,所述将执行语义分析的第二长短时记忆网络与所述第一长短时记忆网络进行通信,以对所述图片内容信息和所述问题信息进行推理结合,得到并输出所述待处理视觉问答的答案,包括:

将所述问题信息和各所述区域特征数据对应的内容信息同时输入执行语义分析的第二长短时记忆网络;其中,所述第二长短时记忆网络每个时间戳的输入包括:第一长短时记忆网络的隐含层输出、所述第二长短时记忆网络上一个时间戳的输出以及所述问题信息中的一个词向量;

基于所述第二长短时记忆网络输出针对问题的答案,并将计算结果输出到softmax层,选择概率最大的词向量作为所述待处理视觉问答的答案进行输出。

6.一种基于多重注意力的视觉问答系统,包括:

信息获取模块,其配置成获取待处理视觉问答中的图片信息以及与所述图片信息对应

的问题信息；

图片特征提取模块，其配置成将提取所述图片信息中的图片特征数据；

图片内容获取模块，其配置成将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络，通过分配注意力权重获取与所述图片特征数据对应的图片内容信息；

答案输出模块，其配置成将执行语义分析的第二长短时记忆网络与所述第一长短时记忆网络进行通信，以对所述图片内容信息和所述问题信息进行推理结合，得到并输出所述待处理视觉问答的答案。

7.根据权利要求6所述的系统，其特征在于，所述图片特征提取模块，其还配置成：

使用区域卷积神经网络R-CNN提取所述图片信息中的至少一个特征区域；

通过ResNet提取每个所述特征区域中的至少一个区域特征信息；

对于每个特征区域，根据重叠度IoU对该特征区域中的区域特征信息进行特征筛选，并将筛选后的区域特征信息作平均池化，进而得到每个特征区域的区域特征数据。

8.根据权利要求6所述的系统，其特征在于，还包括：

预训练模块，其配置成基于预设数据集对所述R-CNN和/或ResNet进行预训练；将所述预设数据集中所包含图片的每个区域特征与代表真实类别的向量融合；将融合后的向量传输至所述R-CNN和/或ResNet的全连接层输出做属性类别和非属性类别的softmax分类。

9.根据权利要求6所述的系统，其特征在于，所述图片内容获取模块，其还配置成：

将所述图片特征数据输入基于注意力机制的第一长短时记忆网络；其中，在所述第一长短时记忆网络的每个时间戳上，其输入包括：上一个时间戳的第二长短时记忆网络的输出、各区域特征数据和上一个时间戳第一长短时记忆网络的输出；其输出包括：对每个所述区域特征数据分配注意力权重；

基于每个所述区域特征数据注意力权重获取与每个所述区域特征数据对应的内容信息。

10.根据权利要求6所述的系统，其特征在于，所述答案输出模块，其还配置成：

将所述问题信息和各所述区域特征数据对应的内容信息同时输入执行语义分析的第二长短时记忆网络；其中，所述第二长短时记忆网络每个时间戳的输入包括：第一长短时记忆网络的隐含层输出、所述第二长短时记忆网络上一个时间戳的输出以及所述问题信息中的一个词向量；

基于所述第二长短时记忆网络输出针对问题的答案，并将计算结果输出到softmax层，选择概率最大的词向量作为所述待处理视觉问答的答案进行输出。

一种基于多重注意力的视觉问答方法及系统

技术领域

[0001] 本申请涉及视觉问答领域,特别是涉及一种基于多重注意力的视觉问答方法及系统。

背景技术

[0002] 视觉问答是一种涉及计算机视觉和自然语言处理的学习任务,就是让计算机学习输入的图片和问题输出一个符合自然语言规则且内容符合逻辑的答案,它根据问题的不同仅聚焦与图片中某一部分的对象,并且某些问题需要一定的常识推理才能得到答案,所以,视觉问答相比于一般的看图说话在对图像的语义理解上要求更高,也面对着更大的挑战。

[0003] 目前视觉问答领域现有模型有Deeper LSTM Q+norm I模型和VIS+LSTM模型等,但除了在回答单一答案的简单问题上有较高的准确率外,其他方面模型的准确率普遍偏低,结构还相对简单,答案的内容和形式比较单一,对于稍复杂的需要更多先验知识进行简单推理的问题无法做出正确的回答。

发明内容

[0004] 本申请的目的在于克服上述问题或者至少部分地解决或缓减解决上述问题。

[0005] 根据本申请的一个方面,提供了一种基于多重注意力的视觉问答方法,包括:

[0006] 获取待处理视觉问答中的图片信息以及与所述图片信息对应的问题信息;

[0007] 提取所述图片信息中的图片特征数据;

[0008] 将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络,通过分配注意力权重获取与所述图片特征数据对应的图片内容信息;

[0009] 将执行语义分析的第二长短时记忆网络与所述第一长短时记忆网络进行通信,以对所述图片内容信息和所述问题信息进行推理结合,得到并输出所述待处理视觉问答的答案。

[0010] 可选地,所述提取所述图片信息中的图片特征数据,包括:

[0011] 使用区域卷积神经网络R-CNN提取所述图片信息中的至少一个特征区域;

[0012] 通过ResNet提取每个所述特征区域中的至少一个区域特征信息;

[0013] 对于每个特征区域,根据重叠度IoU对该特征区域中的区域特征信息进行特征筛选,并将筛选后的区域特征信息作平均池化,进而得到每个特征区域的区域特征数据。

[0014] 可选地,所述提取所述图片信息中的图片特征数据之前,还包括:

[0015] 基于预设数据集对所述R-CNN和/或ResNet进行预训练;

[0016] 将所述预设数据集中所包含图片的每个区域特征与代表真实类别的向量融合;

[0017] 将融合后的向量传输至所述R-CNN和/或ResNet的全连接层输出做属性类别和非属性类别的softmax分类。

[0018] 可选地,所述将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络,通过分配注意力权重获取与所述图片特征数据对应的图片内容信息,

包括：

[0019] 将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络；其中，在所述第一长短时记忆网络的每个时间戳上，其输入包括：上一个时间戳的第二长短时记忆网络的输出、各区域特征数据和上一个时间戳第一长短时记忆网络的输出；其输出包括：对每个所述区域特征数据分配注意力权重；

[0020] 基于每个所述区域特征数据注意力权重获取与每个所述区域特征数据对应的内容信息。

[0021] 可选地，所述将执行语义分析的第二长短时记忆网络与所述第一长短时记忆网络进行通信，以对所述图片内容信息和所述问题信息进行推理结合，得到并输出所述待处理视觉问答的答案，包括：

[0022] 将所述问题信息和各所述区域特征数据对应的内容信息同时输入执行语义分析的第二长短时记忆网络；其中，所述第二长短时记忆网络每个时间戳的输入包括：第一长短时记忆网络的隐含层输出、所述第二长短时记忆网络上一个时间戳的输出以及所述问题信息中的一个词向量；

[0023] 基于所述第二长短时记忆网络输出针对问题的答案，并将计算结果输出到softmax层，选择概率最大的词向量作为所述待处理视觉问答的答案进行输出。

[0024] 根据本申请的另一个方面，提供了一种基于多重注意力的视觉问答系统，包括：

[0025] 信息获取模块，其配置成获取待处理视觉问答中的图片信息以及与所述图片信息对应的问题信息；

[0026] 图片特征提取模块，其配置成将提取所述图片信息中的图片特征数据；

[0027] 图片内容获取模块，其配置成将所述图片特征数据和所述问题信息同时输入基于注意力机制的第一长短时记忆网络，通过分配注意力权重获取与所述图片特征数据对应的图片内容信息；

[0028] 答案输出模块，其配置成将执行语义分析的第二长短时记忆网络与所述第一长短时记忆网络进行通信，以对所述图片内容信息和所述问题信息进行推理结合，得到并输出所述待处理视觉问答的答案。

[0029] 可选地，所述图片特征提取模块，其还配置成：

[0030] 使用区域卷积神经网络R-CNN提取所述图片信息中的至少一个特征区域；

[0031] 通过ResNet提取每个所述特征区域中的至少一个区域特征信息；

[0032] 对于每个特征区域，根据重叠度IoU对该特征区域中的区域特征信息进行特征筛选，并将筛选后的区域特征信息作平均池化，进而得到每个特征区域的区域特征数据。

[0033] 可选地，所述系统还包括：

[0034] 预训练模块，其配置成基于预设数据集对所述R-CNN和/或ResNet进行预训练；将所述预设数据集中所包含图片的每个区域特征与代表真实类别的向量融合；将融合后的向量传输至所述R-CNN和/或ResNet的全连接层输出做属性类别和非属性类别的softmax分类。

[0035] 可选地，所述图片内容获取模块，其还配置成：

[0036] 将所述图片特征数据输入基于注意力机制的第一长短时记忆网络；其中，在所述第一长短时记忆网络的每个时间戳上，其输入包括：上一个时间戳的第二长短时记忆网络

的输出、各区域特征数据和上一个时间戳第一长短时记忆网络的输出;其输出包括:对每个所述区域特征数据分配注意力权重;

[0037] 基于每个所述区域特征数据注意力权重获取与每个所述区域特征数据对应的内容信息。

[0038] 可选地,所述答案输出模块,其还配置成:

[0039] 将所述问题信息和各所述区域特征数据对应的内容信息同时输入执行语义分析的第二长短时记忆网络;其中,所述第二长短时记忆网络每个时间戳的输入包括:第一长短时记忆网络的隐含层输出、所述第二长短时记忆网络上一个时间戳的输出以及所述问题信息中的一个词向量;

[0040] 基于所述第二长短时记忆网络输出针对问题的答案,并将计算结果输出到softmax层,选择概率最大的词向量作为所述待处理视觉问答的答案进行输出。

[0041] 本申请提供了一种基于多重注意力的视觉问答方法及系统,在本申请提供的方法中,先获取待处理视觉问答中的图片信息及对应的问题信息,提取图片信息中的图片特征数据,然后将图片特征数据和问题信息同时输入基于注意力机制的第一长短时记忆网络,通过分配注意力权重获取图片内容信息,再通过两个双向长短时记忆网络完成问题和图片的结合,输出待处理视觉问答的答案。

[0042] 基于本申请提供的基于多重注意力的视觉问答方法及系统,在R-CNN网络的计算过程中增加一些记忆模块以提高模型在训练过程中的知识源,产生更多样化且合理的答案,端到端的记忆网络实现提取问题信息的同时融合与问题相关的先验知识,提高问题回答上的整体准确率。

[0043] 根据下文结合附图对本申请的具体实施例的详细描述,本领域技术人员将会更加明了本申请的上述以及其他目的、优点和特征。

附图说明

[0044] 后文将参照附图以示例性而非限制性的方式详细描述本申请的一些具体实施例。附图中相同的附图标记标示了相同或类似的部件或部分。本领域技术人员应该理解,这些附图未必是按比例绘制的。附图中:

[0045] 图1是根据本申请实施例的基于多重注意力的视觉问答方法流程示意图;

[0046] 图2是根据本申请实施例的双向LSTM工作流程示意图;

[0047] 图3是根据本申请实施例的视觉问答中图片信息示意图;

[0048] 图4是根据本申请实施例的基于多重注意力的视觉问答系统结构示意图。

[0049] 图5是根据本申请优选实施例的基于多重注意力的视觉问答系统结构示意图;

[0050] 图6是根据本申请实施例的计算设备示意图;

[0051] 图7是根据本社情实施例的计算机可读存储介质示意图。

具体实施方式

[0052] 图1是根据本申请实施例的基于多重注意力的视觉问答方法流程示意图。参见图1所知,本申请实施例提供的基于多重注意力的视觉问答方法可以包括:

[0053] 步骤S101:获取待处理视觉问答中的图片信息以及与图片信息对应的问题信息;

[0054] 步骤S102:提取图片信息中的图片特征数据;

[0055] 步骤S103:将上述图片特征数据和问题信息同时输入基于注意力机制的第一长短时记忆网络,通过分配注意力权重获取与图片特征数据对应的图片内容信息;

[0056] 步骤S104:将执行语义分析的第二长短时记忆网络与第一长短时记忆网络进行通信,以对图片内容信息和问题信息进行推理结合,得到并输出待处理视觉问答的答案。

[0057] 本申请实施例提供了一种基于多重注意力的视觉问答方法,在本申请实施例提供的方法中,先获取待处理视觉问答中的图片信息及对应的问题信息,提取图片信息中的图片特征数据,然后将图片特征数据和问题信息同时输入基于注意力机制的第一长短时记忆网络,通过分配注意力权重获取图片内容信息,再通过两个双向长短时记忆网络完成问题和图片的结合,输出待处理视觉问答的答案。

[0058] 长短时记忆网络(Long Short-Term Memory,缩写为LSTM),是一种时间递归神经网络,适合于处理和预测时间序列中间隔和延迟相对较长的重要事件,基于LSTM的系统可以学习翻译语言、控制机器人、图像分析、文档摘要、语音识别、图像识别、手写识别、控制聊天机器人、预测疾病、点击率和股票、合成音乐等等任务。

[0059] 对于传统的视觉问答模型Deeper LSTM Q+norm I模型来讲,其中I表示提取后的图片特征,norm I表示由CNN提取的1024维像素语义向量做L2归一化处理。由CNN提取图像语义信息,再通过LSTM拿到问题中包含的文本语义信息,融合两个网络的数据,让模型学习到问题的含义,最后送入一个多层MLP作为Softmax输出层产生答案输出。输入包含了一个室外场景中的2匹马和2个人的图像,这些数据通过不含分类层的CNN处理,问题部分的语义则通过一个RNN网络按照问题单词的输入顺序挨个提取问题信息,然后将两个压缩后的信息融合,把处理后的数据送入MLP中产生结果输出(比如当前问题是计数(count object)问题)。该模型使用2层LSTM编码问题并用VGGNet模型给图像划分区域,接着把图像特征作L2归一化。之后将图像和问题特征变换到同一个特征空间,通过点乘的方式将信息融合,融合后的信息送入一个以Softmax为分类器的三层MLP中产生答案输出。模型在训练的过程中,pre-CNN,只有LSTM层和最后的分类网络参加训练。

[0060] VIS+LSTM模型,其基本结构是首先使用CNN抽取图片信息,完成之后接LSTM产生预测结果。但其考虑到由于没有一个完好的评价答案句子精确度的标准,因此他们将注意力集中在有限的领域问题,这些问题可以用一个单词作为视觉问答的答案,这样就可以把视觉问答视为一个多分类问题,从而可以利用现有的精确度评价标准度量答案。

[0061] 上述所提及各模型的整体正确率并不高,并且答案的内容和形式比较单一。

[0062] 在本实施例中,基于注意力机制的第一长短时记忆网络(以下可简称为注意力LSTM),主要是在LSTM中训练一个模型来对输入序列进行选择性的学习并且在模型输出时会选择性地专注考虑输入中的对应相关的信息。执行语义分析的第二长短时记忆网络(以下可简称为语义LSTM),主要是在LSTM中挖掘与学习文本、图片等的深层次概念,可避免梯度消失问题。其中,注意力LSTM和语义LSTM均优选为双向LSTM,并且可以在二者之间进行通信,协同完成视觉问答中图片和问题的结合,以准确并快速输出问题的答案。

[0063] 通常情况下,视觉问答中为先输入图片和自然语言问题,进而根据自然语言问题聚焦图片信息,从而产生一条人类语言作为答案输出。因此,在解决视觉问答时,需先执行上述步骤S101,获取视觉问答中的图片信息和问题信息。

[0064] 进一步地，即可执行步骤S102，对其中的图片信息进行分析以提取特征数据。可选地，其可以包括：使用R-CNN提取图片信息中的至少一个特征区域；再通过ResNet提取每个特征区域中的至少一个区域特征信息；然后对于每个特征区域，根据重叠度IoU对该特征区域中的区域特征信息进行特征筛选，并将筛选后的区域特征信息作平均池化，进而得到每个特征区域的特征数据。

[0065] R-CNN的作用就是为了找到图片信息中感兴趣的特征区域，并进一步利用ResNet在提取出的特征区域中提取区域特征信息。

[0066] R-CNN的全称是Region-CNN，是第一个成功将深度学习应用到目标检测上的算法。R-CNN基于卷积神经网络(CNN)、线性回归、和支持向量机(SVM)等算法，实现目标检测技术。

[0067] ResNet又叫深度残差网络，与普通网络不同的地方就是引入了跳跃连接，这可以使上一个残差块的信息没有阻碍的流入到下一个残差块，提高了信息流通，并且也避免了由与网络过深所引起的消失梯度问题和退化问题。

[0068] 在提取出各特征区域后的区域特征信息之后，即可根据重叠度IoU对各特征区域中的区域特征信息进行特征筛选。其中，IoU全称Intersection overUnion，是一种测量在特定数据集中检测相应物体准确度的一个标准。通过基于IoU的值与预设阈值作比较，对区域内的特征数据进行筛选，以进一步缩小图片信息。最后对筛选后的特征用平均池化的卷积特征来表示，获取每个特征区域的区域特征数据，另外，可对其进行拼接以获取问答系统中图片信息的特征拼接图，并作为R-CNN的输出结果。

[0069] 在本发明一可选实施例中，在利用R-CNN和ResNet进行特征提取之前，可先对R-CNN和ResNet进行预训练，将感兴趣区域和可能的类别融合。具体可以包括：基于预设数据集对R-CNN和/或ResNet进行预训练；将预设数据集中所包含图片的每个区域特征与代表真实类别的向量融合；将融合后的向量传输至R-CNN和/或ResNet的全连接层输出做属性类别和非属性类别的softmax分类。

[0070] 本实施例中，通过对R-CNN和ResNet进行预训练，可以优化模型的参数，让模型可以找到感兴趣的区域。其中，预设数据集可以优选为COCO，全称是Common Objects in COntext，是一个可以用来进行图像识别的数据集。MS COCO数据集中的图像分为训练、验证和测试集。COCO通过在搜索引擎上搜索80个对象类别和各种场景类型来收集图像。COCO数据集现在有3种标注类型：object instances(目标实例)，objectkeypoints(目标上的关键点)，和image captions(看图说话)，使用JSON文件存储。相比较现有模型如Deeper LSTM Q +norm I模型、VIS+LSTM模型等，问题回答上有更高的准确率。

[0071] 在得到图像信息的特征数据之后，就可以通过两个双向LSTM完成问题和图片的结合，并输出答案。也就是说，在本实施例所提供方法中，可先对图片的数据做预处理，先把图片中对应区域与对应类别做映射，一个注意力机制的LSTM，把单词和图片中的特定区域融合，分析特定区域的内容，最后把注意力机制的LSTM输出结果拿到语义LSTM中，把单词做推理结合生成问题答案。

[0072] 参见上述步骤S103，将图片特征数据和问题信息同时输入基于注意力机制的LSTM，通过分配注意力权重获取与图片特征数据对应的图片内容信息。其可以包括，将图片特征数据和问题信息同时输入基于注意力机制的LSTM；其中，在基于注意力机制的LSTM的每个时间戳上，其输入包括：上一个时间戳的语义LSTM的输出、各区域特征数据和上一个时

间戳基于注意力机制的LSTM的输出;其输出包括:对每个所述区域特征数据分配注意力权重;然后基于每个所述区域特征数据注意力权重获取与每个区域特征数据对应的内容信息。

[0073] 在每个时间戳上,结合上一时间戳的两个LSTM的输出和提取出来的图片特征数据,注意力LSTM的每一个cell都会有输出,并且通过时间戳的推移,会给所有的区域特征分配不同的注意力权重,这个是需要学习的参数,将各时间戳上的输出和注意力权重相结合,输出一个供语义LSTM处理的数据。

[0074] 然后执行步骤S104,将语义LSTM与注意力LSTM进行通信,以对图片内容信息和问题信息进行推理结合,得到并输出待处理视觉问答的答案,可以包括:将问题信息和各区域特征数据对应的内容信息同时输入语义LSTM;其中,语义LSTM每个时间戳的输入包括:注意力LSTM的隐含层输出、语义LSTM上一个时间戳的输出以及问题信息中的一个词向量;基于语义LSTM输出针对问题的答案,并将计算结果输出到softmax层,选择概率最大的词向量作为待处理视觉问答的答案进行输出。

[0075] 其中,注意力LSTM的隐含层输出,包含了注意力权重,是注意力LSTM当前时间戳输出的特征数据和注意力权重相结合的计算结果。

[0076] 上文提及的问题的词向量,以检测到“?”表示句子结尾,它的本质是一个词向量矩阵,对应到一个单词的one-hot编码,词向量是随机生成的,没有经过预训练。

[0077] 上述步骤S103和S104涉及到的注意力LSTM和语义LSTM是双向作用的,如图2所示,假设时间序列为t＝{1,2,……,n}两个双向LSTM工作流程可以包括:

[0078] S1-1,t_1时刻,将区域特征数据输入t_1时刻的注意力LSTM,再将t_1时刻注意力LSTM的输出和问题信息中第一个词向量输入t_1时刻的语义LSTM;

[0079] S1-2,t_2时刻,将t_1时刻的两个LSTM的输出和区域特征数据输入t_2时刻的注意力LSTM,然后将t_2时刻基于注意力LSTM的输出、t_1时刻的语义LSTM的输出和问题信息中第二个词向量输入t_2时刻的语义LSTM;

[0080] ……

[0081] S1-n,t_n时刻,将t_{n-1}时刻的两个LSTM的输出和区域特征数据输入t_n时刻的注意力LSTM,然后将t_n时刻的注意力LSTM的输出、t_{n-1}时刻的语义LSTM的输出和问题信息中第n个词向量(词向量以?作为结束,可能比n时刻短)输入t_n时刻的语义LSTM。

[0082] 最终,由t_n时刻的语义LSTM输出问题的答案。

[0083] 举例来讲,假设视觉问答中的图片信息如图3所示,问题信息为“床上的是什么?”,则基于本实施例提供的视觉问答方法可以包括:

[0084] S2-1,先获取此视觉问答中的图片信息和问题信息;

[0085] S2-2,通过R-CNN和ResNet提取图片中的特征数据,其中,通过R-CNN提取图片中感兴趣的多个特征区域,如图片区域1桌子,图片区域2床;再通过ResNet在每个特征区域中提取多个区域特征信息;然后根据重叠度IoU对每个特征区域中的特征进行筛选,再对每个区域内筛选后的特征作平均池化,进而获取图片特征数据。其中,R-CNN和ResNet是经过预训练的;

[0086] S2-3,将图片特征数据和问题信息同时输入注意力LSTM,获取图片内容信息,如区域1中有电脑、台灯、书架、书等,区域2中有书、药等。进一步地,还会为各个区域中所包括的

特征数据分配权重，如区域2中的书占较重比例，药占较小比例等等；将注意力LSTM的输出和注意力权重相结合，输出最终的图片内容信息，如区域2中通过分配不同的权重之后获取图片内容信息为书

[0087] S2-4，将语义LSTM与注意力LSTM双向通信，结合图片内容信息和问题信息推理结合，将语义LSTM最后一步的结果输出到softmax层，选择概率最大词向量作为答案进行输出。问题指向区域2，输出答案books。

[0088] 基于同一发明构思，如图4所示，本申请实施例还提供了一种基于多重注意力的视觉问答系统400，包括：

[0089] 信息获取模块410，其配置成获取待处理视觉问答中的图片信息以及与图片信息对应的问题信息；

[0090] 图片特征提取模块420，其配置成将提取图片信息中的图片特征数据；

[0091] 图片内容获取模块430，其配置成将图片特征数据和所述问题信息同时输入基于注意力机制的LSTM，通过分配注意力权重获取与图片特征数据对应的图片内容信息；

[0092] 答案输出模块440，其配置成将语义LSTM与注意力LSTM进行通信，以对图片内容信息和所述问题信息进行推理结合，得到并输出待处理视觉问答的答案。

[0093] 在本发明一可选实施例中，图片特征提取模块420，其还配置成：

[0094] 使用区域卷积神经网络R-CNN提取所述图片信息中的至少一个特征区域；通过ResNet提取每个所述特征区域中的至少一个区域特征信息；对于每个特征区域，根据重叠度IoU对该特征区域中的区域特征信息进行特征筛选，并将筛选后的区域特征信息作平均池化，进而得到每个特征区域的区域特征数据。

[0095] 在本发明一可选实施例中，如图5所示，上述系统还可以包括：

[0096] 预训练模块450，其配置成基于预设数据集对所述R-CNN和/或ResNet进行预训练；将所述预设数据集中所包含图片的每个区域特征与代表真实类别的向量融合；将融合后的向量传输至所述R-CNN和/或ResNet的全连接层输出做属性类别和非属性类别的softmax分类。

[0097] 在本发明一可选实施例中，图片内容获取模块430，其还配置成：

[0098] 将图片特征数据输入基于注意力机制的LSTM；其中，在注意力LSTM的每个时间戳上，其输入包括：上一个时间戳的语义LSTM的输出、各区域特征数据和上一个时间戳注意力LSTM的输出；其输出包括：对每个所述区域特征数据分配注意力权重；

[0099] 基于每个区域特征数据注意力权重获取与每个区域特征数据对应的内容信息。

[0100] 在本发明一可选实施例中，答案输出模块440，其还配置成：

[0101] 将问题信息和各区域特征数据对应的内容信息同时输入语义LSTM；其中，语义LSTM每个时间戳的输入包括：注意力LSTM的隐含层输出、语义LSTM上一个时间戳的输出以及问题信息中的一个词向量；

[0102] 基于语义LSTM输出针对问题的答案，并将计算结果输出到softmax层，选择概率最大的词向量作为所述待处理视觉问答的答案进行输出。

[0103] 本申请实施例结合了计算机视觉(CV)和自然语言处理(NLP)两大领域，提供了一种基于多重注意力的视觉问答方法及系统，在本申请实施例提供的方法中，先获取待处理视觉问答中的图片信息及对应的问题信息，提取图片信息中的图片特征数据，然后将图片

特征数据和问题信息同时输入基于注意力机制的第一长短时记忆网络，通过分配注意力权重获取图片内容信息，再通过两个双向长短时记忆网络完成问题和图片的结合，输出待处理视觉问答的答案。

[0104] 基于本申请实施例提供的基于多重注意力的视觉问答方法及系统，在R-CNN网络的计算过程中增加一些记忆模块以提高模型在训练过程中的知识源，产生更多样化且合理的答案，端到端的记忆网络实现提取问题信息的同时融合与问题相关的先验知识，提高问题回答上的整体准确率。

[0105] 根据下文结合附图对本申请的具体实施例的详细描述，本领域技术人员将会更加明了本申请的上述以及其他目的、优点和特征。

[0106] 本申请实施例还提供了一种计算设备，参照图6，该计算设备包括存储器620、处理器610和存储在所述存储器620内并能由所述处理器610运行的计算机程序，该计算机程序存储于存储器620中的用于程序代码的空间630，该计算机程序在由处理器610执行时实现用于执行任一项根据本发明的方法步骤631。

[0107] 本申请实施例还提供了一种计算机可读存储介质。参照图7，该计算机可读存储介质包括用于程序代码的存储单元，该存储单元设置有用于执行根据本发明的方法步骤的程序631′，该程序被处理器执行。

[0108] 本申请实施例还提供了一种包含指令的计算机程序产品。当该计算机程序产品在计算机上运行时，使得计算机执行根据本发明的方法步骤。

[0109] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、获取其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘Solid State Disk(SSD))等。

[0110] 专业人员应该还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，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。专业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申请的范围。

[0111] 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步骤是可以通过程序来指令处理器完成，所述的程序可以存储于计算机可读存储介质中，所述存储介质是非短暂性(英文:non-transitory)介质，例如随机存取存储器，只读存储器，快闪存储

器,硬盘,固态硬盘,磁带(英文:magnetic tape),软盘(英文:floppy disk),光盘(英文:optical disc)及其任意组合。

[0112] 以上所述,仅为本申请较佳的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到的变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应该以权利要求的保护范围为准。

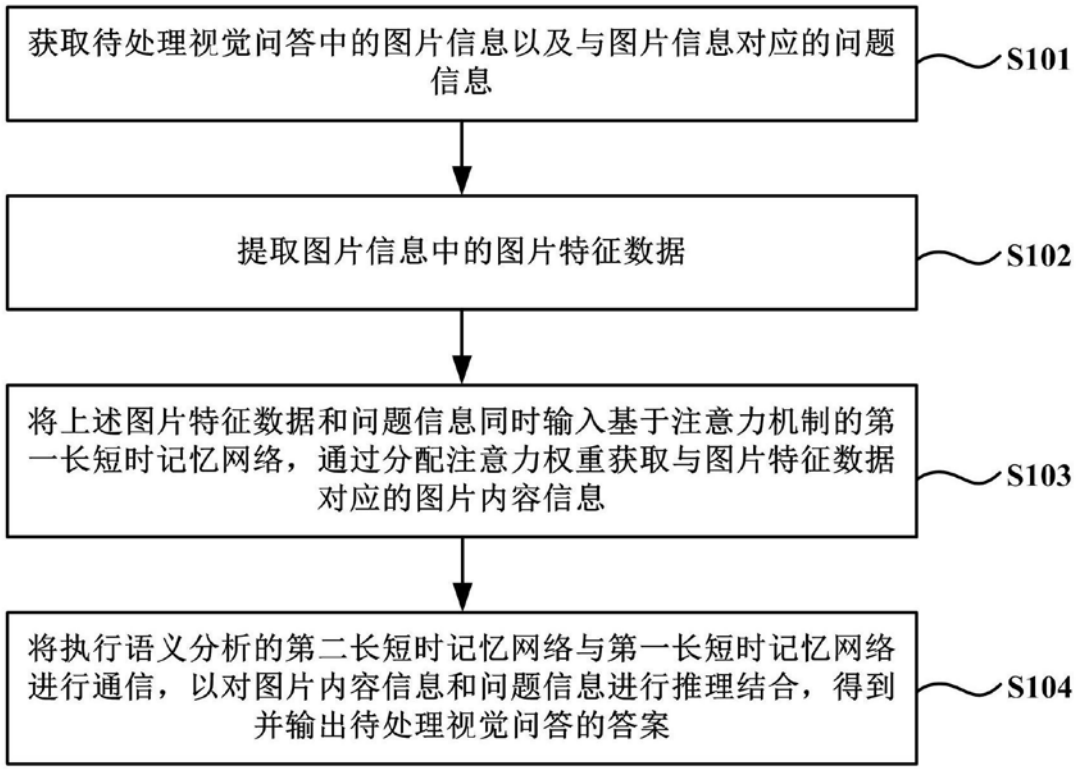

获取待处理视觉问答中的图片信息以及与图片信息对应的问题信息
S101
提取图片信息中的图片特征数据
S102
将上述图片特征数据和问题信息同时输入基于注意力机制的第一长短时记忆网络，通过分配注意力权重获取与图片特征数据对应的图片内容信息
S103
将执行语义分析的第二长短时记忆网络与第一长短时记忆网络进行通信，以对图片内容信息和问题信息进行推理结合，得到并输出待处理视觉问答的答案
S104

图1

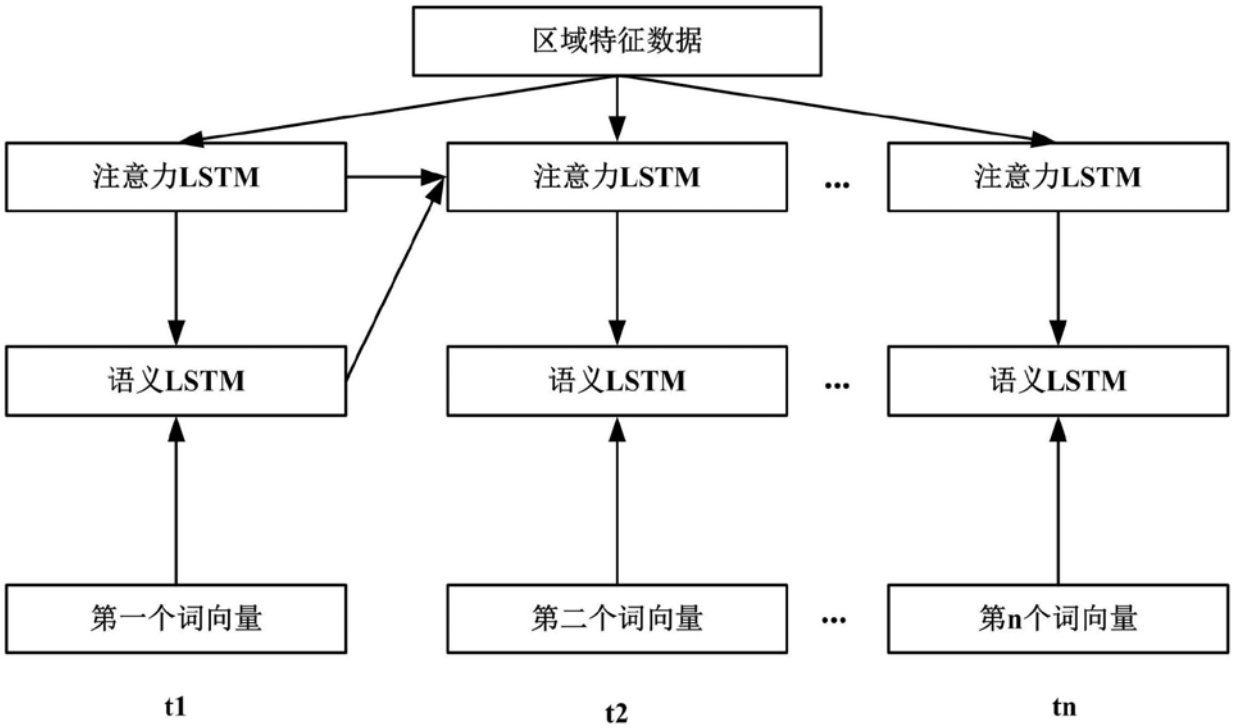

区域特征数据
注意力LSTM
注意力LSTM
...
注意力LSTM
语义LSTM
语义LSTM
...
语义LSTM
第一个词向量
第二个词向量
...
第n个词向量
t1
t2
tn

图2

图3

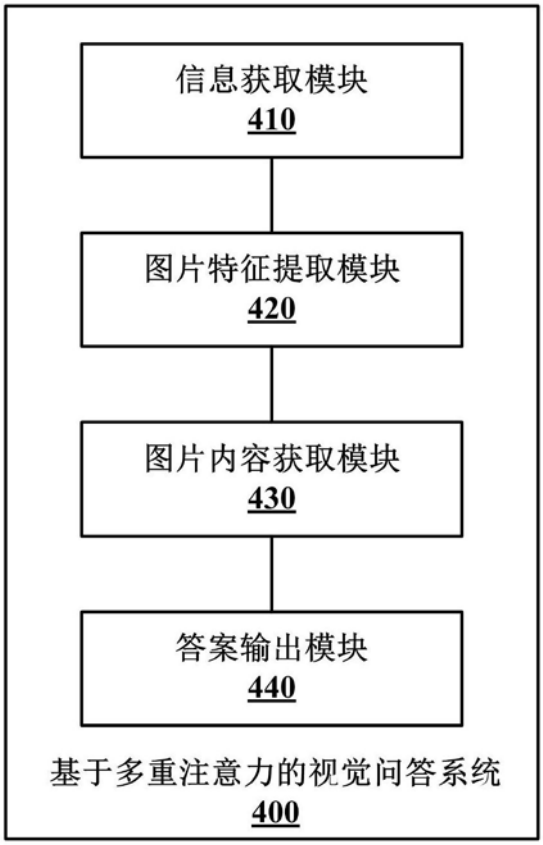

信息获取模块
410
图片特征提取模块
420
图片内容获取模块
430
答案输出模块
440
基于多重注意力的视觉问答系统
400

图4

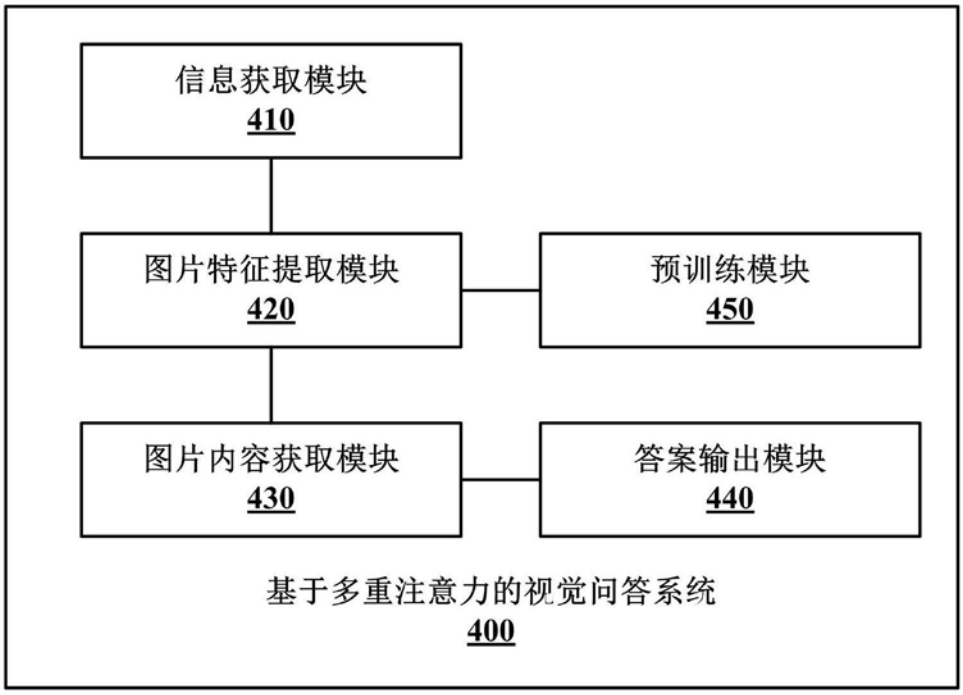

信息获取模块
410
图片特征提取模块
420
预训练模块
450
图片内容获取模块
430
答案输出模块
440
基于多重注意力的视觉问答系统
400

图5

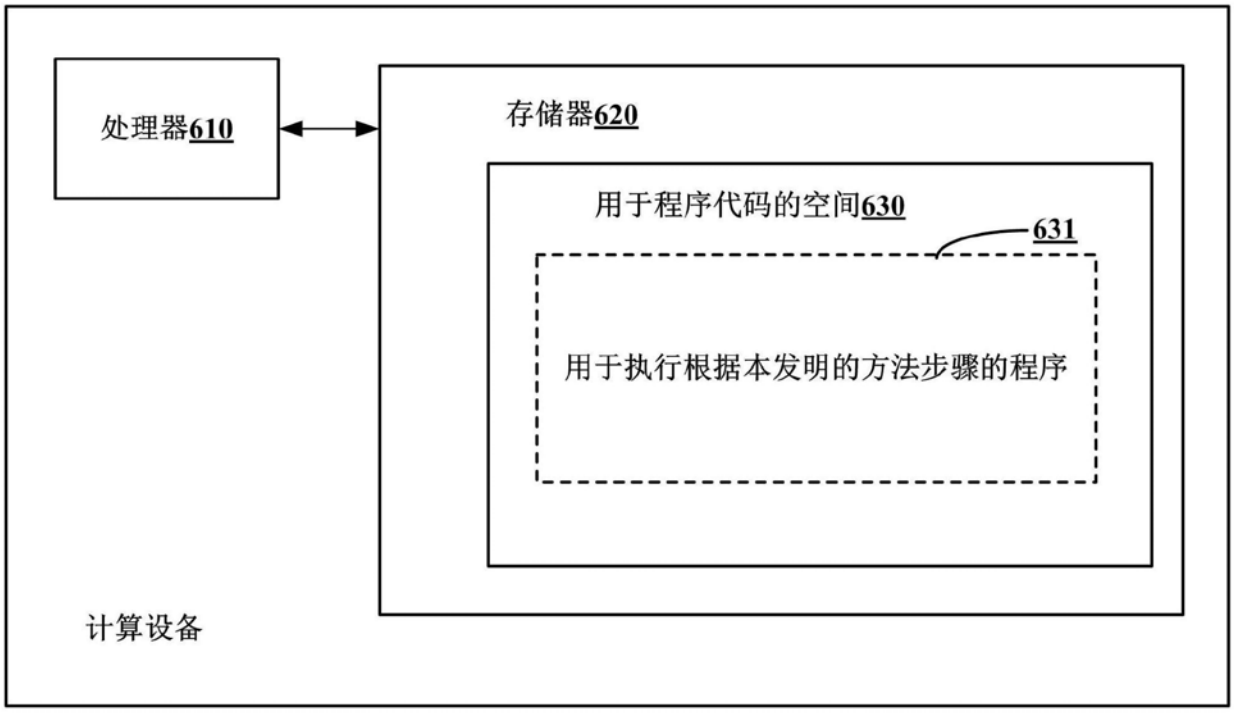

处理器610
存储器620
用于程序代码的空间630
631
用于执行根据本发明的方法步骤的程序
计算设备

图6

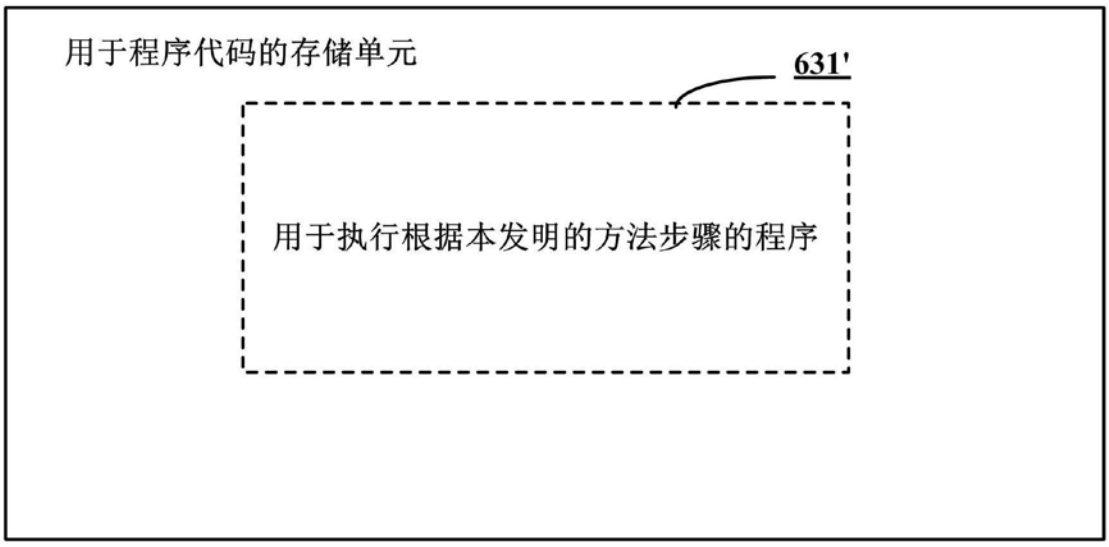


图7