(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 110135441 B
(45)授权公告日 2020.03.03

(21)申请号 201910414090.1

(22)申请日 2019.05.17

(65)同一申请的已公布的文献号
申请公布号 CN 110135441 A

(43)申请公布日 2019.08.16

(73)专利权人 北京邮电大学
地址 100876 北京市海淀区西土城路10号

(72)发明人 李睿凡 梁昊雨 石祎晖 冯方向 张光卫 王小捷

(74)专利代理机构 北京柏杉松知识产权代理事务所(普通合伙) 11413
代理人 丁芸 马敬

(51)Int.Cl.
G06K 9/46(2006.01)
G06K 9/62(2006.01)

(56)对比文件
CN 109002852 A,2018.12.14,
CN 109711464 A,2019.05.03,
US 2018373979 A1,2018.12.27,
Krause J."A Hierarchical Approach for Generating Descriptive Image Paragraphs".《arXiv》.2017,第1-9页.

审查员 张健

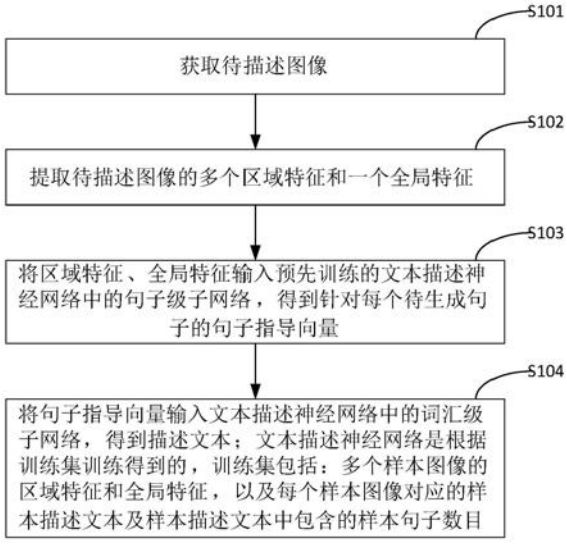
权利要求书3页 说明书10页 附图3页

(54)发明名称

一种图像的文本描述方法及装置

(57)摘要

本发明实施例提供了一种图像的文本描述方法及装置,方法包括:获取待描述图像,提取待描述图像的多个区域特征和一个全局特征;将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络,得到针对每个待生成句子的句子指导向量;将句子指导向量输入文本描述神经网络中的词汇子网络,得到描述文本;由于采用句子级子网络和词汇级子网络的分层结构,能够捕捉段落中句子之间的连贯性,提高了生成的文本段落中句子之间的连贯性,此外,相较于现有的基于循环神经网络的方案,降低了训练过程的计算复杂度。



CN 110135441 B

1.一种图像的文本描述方法,其特征在于,所述方法包括:

获取待描述图像;

提取所述待描述图像的多个区域特征和一个全局特征;

将所述区域特征、所述全局特征输入预先训练的文本描述神经网络中的句子级子网络,得到针对每个待生成句子的句子指导向量;

将所述句子指导向量输入所述文本描述神经网络中的词汇级子网络,得到描述文本;所述文本描述神经网络是根据训练集训练得到的,所述训练集包括:多个样本图像的区域特征和全局特征,以及每个样本图像对应的样本描述文本及所述样本描述文本中包含的样本句子数目;

所述句子级子网络中包含句子嵌入层,门控卷积层和区域感知层,所述将所述区域特征、所述全局特征输入预先训练的文本描述神经网络中的句子级子网络,得到针对每个待生成句子的句子指导向量的步骤,包括:

获取当前已生成的每个句子文本;

将所述每个句子文本输入所述句子嵌入层,得到所述每个句子文本的句子嵌入向量;

将所述每个句子嵌入向量均与所述全局特征的向量拼接,得到多个拼接向量;

将每个所述拼接向量输入所述门控卷积层,得到针对当前待生成句子的隐藏向量;

将所述区域特征的向量和所述隐藏向量输入所述区域感知层,得到针对当前待生成句子的句子指导向量。

2.根据权利要求1所述的方法,其特征在于,所述门控卷积层根据所述每个拼接向量,得到针对当前待生成句子的隐藏向量,包括:

基于每个所述拼接向量进行卷积运算,得到针对当前待生成句子的综合语义向量以及门控向量;

基于所述综合语义向量,以及门控向量,进行语义筛选运算,得到针对当前待生成句子的隐藏向量。

3.根据权利要求2所述的方法,其特征在于,所述区域感知层根据所述区域特征的向量,和所述隐藏向量得到针对当前待生成句子的句子指导向量,包括:

基于所述区域特征的向量,以及所述隐藏向量,计算针对每个所述区域特征的向量的权重;

基于所述权重,计算针对当前待生成句子的加权区域特征向量;

基于所述加权区域特征向量,以及所述隐藏向量,计算针对当前待生成句子的句子指导向量。

4.根据权利要求1所述的方法,其特征在于,所述文本描述神经网络采用如下步骤训练获得:

获取预设的神经网络模型和所述训练集;

将所述样本图像的区域特征和全局特征输入所述神经网络模型,得到描述文本以及所述描述文本中包含的句子数目;

基于得到的描述文本以及句子数目,和所述训练集中包含的所述样本描述文本及样本句子数目,确定损失值;

根据所述损失值确定所述神经网络模型是否收敛;

若否，则调整所述神经网络模型中的参数值，并返回将所述样本图像的区域特征和全局特征输入所述神经网络模型，得到描述文本以及所述描述文本中包含的句子数目的步骤；

若是，则将当前的神经网络模型确定为文本描述神经网络。

5.一种图像的文本描述装置，其特征在于，所述装置包括：

获取模块，用于获取待描述图像；

提取模块，用于提取所述待描述图像的多个区域特征和一个全局特征；

第一输入模块，用于将所述区域特征、所述全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量；

第二输入模块，用于将所述句子指导向量输入所述文本描述神经网络中的词汇级子网络，得到描述文本；所述文本描述神经网络是根据训练集训练得到的，所述训练集包括：多个样本图像的区域特征和全局特征，以及每个样本图像对应的样本描述文本及所述样本描述文本中包含的样本句子数目；

所述句子级子网络中包含句子嵌入层，门控卷积层和区域感知层，所述第一输入模块，具体用于：

获取当前已生成的每个句子文本；

将所述每个句子文本输入所述句子嵌入层，得到所述每个句子文本的句子嵌入向量；

将所述每个句子嵌入向量均与所述全局特征的向量拼接，得到多个拼接向量；

将每个所述拼接向量输入所述门控卷积层，得到针对当前待生成句子的隐藏向量；

将所述区域特征的向量和所述隐藏向量输入所述区域感知层，得到针对当前待生成句子的句子指导向量。

6.根据权利要求5所述的装置，其特征在于，所述门控卷积层，具体用于：

基于每个所述拼接向量进行卷积运算，得到针对当前待生成句子的综合语义向量以门控向量；

基于所述综合语义向量，以及门控向量，进行语义筛选运算，得到针对当前待生成句子的隐藏向量；

所述区域感知层，具体用于：

基于所述区域特征的向量，以及所述隐藏向量，计算针对每个所述区域特征的向量的权重；

基于所述权重，计算针对当前待生成句子的加权区域特征向量；

基于所述加权区域特征向量，以及所述隐藏向量，计算针对当前待生成句子的句子指导向量。

7.根据权利要求5所述的装置，其特征在于，所述装置还包括训练模块，所述训练模块，具体用于：

获取预设的神经网络模型和所述训练集；

将所述样本图像的区域特征和全局特征输入所述神经网络模型，得到描述文本以及所述描述文本中包含的句子数目；

基于得到的描述文本以及句子数目，和所述训练集中包含的所述样本描述文本及样本句子数目，确定损失值；

根据所述损失值确定所述神经网络模型是否收敛；

若否，则调整所述神经网络模型中的参数值，并返回将所述样本图像的区域特征和全局特征输入所述神经网络模型，得到描述文本以及所述描述文本中包含的句子数目的步骤；

若是，则将当前的神经网络模型确定为文本描述神经网络。

8.一种电子设备，其特征在于，包括处理器、通信接口、存储器和通信总线，其中，处理器，通信接口，存储器通过通信总线完成相互间的通信；

存储器，用于存放计算机程序；

处理器，用于执行存储器上所存放的程序时，实现权利要求1-4任一所述的方法步骤。

一种图像的文本描述方法及装置

技术领域

[0001] 本发明涉及图像描述技术领域，特别是涉及一种图像的文本描述方法及装置。

背景技术

[0002] 图像描述是指用自然语言描述给定图像的内容，通过借用注意力机制增强的编码器-解码器框架，单句描述任务已经取得了较大进展。然而，一个句子不足以描述一个语义丰富的图像。因此，人们又提出了图像段落描述，即使用连贯的段落来描述图像。

[0003] 现有的图像段落描述是基于RNN(Recurrent Neural Network，循环神经网络)的，这种方式存在以下缺陷：

[0004] 首先，RNN很难记忆长期信息，从而限制了其在语言建模上的性能。使用RNN解码器生成的句子的连贯性较低。其次，由于RNN的序列特点，RNN 解码器的训练算法具有很高的计算复杂度。

发明内容

[0005] 本发明实施例的目的在于提供一种图像的文本描述方法及装置，以实现增强文本描述的连贯性，并降低计算复杂度。具体技术方案如下：

[0006] 为了实现上述目的，本发明实施例提供了一种图像的文本描述方法，方法包括：

[0007] 获取待描述图像；

[0008] 提取所述待描述图像的多个区域特征和一个全局特征；

[0009] 将所述区域特征、所述全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量；

[0010] 将所述句子指导向量输入所述文本描述神经网络中的词汇级子网络，得到描述文本；所述文本描述神经网络是根据训练集训练得到的，所述训练集包括：多个样本图像的区域特征和全局特征，以及每个样本图像对应的样本描述文本及所述样本描述文本中包含的样本句子数目。

[0011] 可选的，所述句子级子网络中包含句子嵌入层，门控卷积层和区域感知层，所述将所述区域特征、所述全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量的步骤，包括：

[0012] 获取当前已生成的每个句子文本；

[0013] 将所述每个句子文本输入所述句子嵌入层，得到所述每个句子文本的句子嵌入向量；

[0014] 将所述每个句子嵌入向量均与所述全局特征的向量拼接，得到多个拼接向量；

[0015] 将每个所述拼接向量输入所述门控卷积层，得到针对当前待生成句子的隐藏向量；

[0016] 将所述区域特征的向量和所述隐藏向量输入所述区域感知层，得到针对当前待生成句子的句子指导向量。

[0017] 可选的，所述门控卷积层根据所述每个拼接向量，得到针对当前待生成句子的隐藏向量，包括：

[0018] 基于每个所述拼接向量进行卷积运算，得到针对当前待生成句子的综合语义向量以及门控向量；

[0019] 基于所述综合语义向量，以及门控向量，进行语义筛选运算，得到针对当前待生成句子的隐藏向量。

[0020] 可选的，所述区域感知层根据所述区域特征的向量，和所述隐藏向量得到针对当前待生成句子的句子指导向量，包括：

[0021] 基于所述区域特征的向量，以及所述隐藏向量，计算针对每个所述区域特征的向量的权重；

[0022] 基于所述权重，计算针对当前待生成句子的加权区域特征向量；

[0023] 基于所述加权区域特征向量，以及所述隐藏向量，计算针对当前待生成句子的句子指导向量。

[0024] 可选的，所述文本描述神经网络采用如下步骤训练获得：

[0025] 获取预设的神经网络模型和所述训练集；

[0026] 将所述样本图像的区域特征和全局特征输入所述神经网络模型，得到描述文本以及所述描述文本中包含的句子数目；

[0027] 基于得到的描述文本以及句子数目，和所述训练集中包含的所述样本描述文本及样本句子数目，确定损失值；

[0028] 根据所述损失值确定所述神经网络模型是否收敛；

[0029] 若否，则调整所述神经网络模型中的参数值，并返回将所述样本图像的区域特征和全局特征输入所述神经网络模型，得到描述文本以及所述描述文本中包含的句子数目的步骤；

[0030] 若是，则将当前的神经网络模型确定为文本描述神经网络。

[0031] 为了实现上述目的，本发明实施例还提供了一种图像的文本描述装置，所述装置包括：

[0032] 获取模块，用于获取待描述图像；

[0033] 提取模块，用于提取所述待描述图像的多个区域特征和一个全局特征；

[0034] 第一输入模块，用于将所述区域特征、所述全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量；

[0035] 第二输入模块，用于将所述句子指导向量输入所述文本描述神经网络中的词汇级子网络，得到描述文本；所述文本描述神经网络是根据训练集训练得到的，所述训练集包括：多个样本图像的区域特征和全局特征，以及每个样本图像对应的样本描述文本及所述样本描述文本中包含的样本句子数目。

[0036] 可选的，所述句子级子网络中包含句子嵌入层，门控卷积层和区域感知层，所述第一输入模块，具体用于：

[0037] 获取当前已生成的每个句子文本；

[0038] 将所述每个句子文本输入所述句子嵌入层，得到所述每个句子文本的句子嵌入向量；

[0039] 将所述每个句子嵌入向量均与所述全局特征的向量拼接，得到多个拼接向量；

[0040] 将每个所述拼接向量输入所述门控卷积层，得到针对当前待生成句子的隐藏向量；

[0041] 将所述区域特征的向量和所述隐藏向量输入所述区域感知层，得到针对当前待生成句子的句子指导向量。

[0042] 所述门控卷积层，具体用于：

[0043] 基于每个所述拼接向量进行卷积运算，得到针对当前待生成句子的综合语义向量以及门控向量；

[0044] 基于所述综合语义向量，以及门控向量，进行语义筛选运算，得到针对当前待生成句子的隐藏向量。

[0045] 所述区域感知层，具体用于：

[0046] 基于所述区域特征的向量，以及所述隐藏向量，计算针对每个所述区域特征的向量的权重；

[0047] 基于所述权重，计算针对当前待生成句子的加权区域特征向量；

[0048] 基于所述加权区域特征向量，以及所述隐藏向量，计算针对当前待生成句子的句子指导向量。

[0049] 可选的，所述装置还包括训练模块，所述训练模块，具体用于：

[0050] 获取预设的神经网络模型和所述训练集；

[0051] 将所述样本图像的区域特征和全局特征输入所述神经网络模型，得到描述文本以及所述描述文本中包含的句子数目；

[0052] 基于得到的描述文本以及句子数目，和所述训练集中包含的所述样本描述文本及样本句子数目，确定损失值；

[0053] 根据所述损失值确定所述神经网络模型是否收敛；

[0054] 若否，则调整所述神经网络模型中的参数值，并返回将所述样本图像的区域特征和全局特征输入所述神经网络模型，得到描述文本以及所述描述文本中包含的句子数目的步骤；

[0055] 若是，则将当前的神经网络模型确定为文本描述神经网络。

[0056] 为了实现上述目的，本发明实施例还提供了一种电子设备，包括处理器、通信接口、存储器和通信总线，其中，处理器，通信接口，存储器通过通信总线完成相互间的通信；

[0057] 存储器，用于存放计算机程序；

[0058] 处理器，用于执行存储器上所存放的程序时，实现上述任一方法步骤。

[0059] 为了解决上述技术问题，本发明实施例还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内存储有计算机程序，所述计算机程序被处理器执行时实现上述任一方法步骤。

[0060] 本发明实施例提供的图像的文本描述方法及装置，能够获取待描述图像，提取待描述图像的多个区域特征和一个全局特征；将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量；将句子指导向量输入文本描述神经网络中的词汇子网络，得到描述文本；由于采用句子级子网络和词汇级子网络的分层结构，能够捕捉段落中句子之间的连贯性，提高了生成的文本段落中句子

之间的连贯性，此外，相较于现有的未进行分级训练的方案，降低了训练过程的计算复杂度。

[0061] 当然，实施本发明的任一产品或方法必不一定需要同时达到以上所述的所有优点。

附图说明

[0062] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0063] 图1为本发明实施例提供的图像的文本描述方法的一种流程图；

[0064] 图2为本发明实施例提供的待描述图像的一种示意图；

[0065] 图3为本发明实施例提供的图像的文本描述方法的一种流程示意图；

[0066] 图4为本发明实施例提供的图像的文本描述方法的另一种流程示意图；

[0067] 图5为本发明实施例提供的图像的文本描述装置的一种结构示意图；

[0068] 图6为本发明实施例提供的电子设备的一种结构示意图。

具体实施方式

[0069] 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，所描述的实施例仅仅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。

[0070] 为了增强文本描述的连贯性，并降低计算复杂度，本发明实施例提供了一种图像的文本描述方法，参见图1，图1为本发明实施例提供的图像的文本描述方法的一种流程图，该方法可以应用于电子设备，方法包括以下步骤：

[0071] S101：获取待描述图像。

[0072] 本发明实施例中，待描述图像可以是包含一定场景信息的图像，举例来讲，参见图2，图2是本发明实施例提供的待描述图像的一种示意图。

[0073] S102：提取待描述图像的多个区域特征和一个全局特征。

[0074] 在本步骤中，可以采用图片特征提取器来提取待描述图像的区域特征和全局特征。

[0075] 一种实施例中，可以采用图像编码器来检测待描述图像中的不同图像区域，并提取每个图像区域的特征，得到一组区域特征$\{v_1, v_2, \cdots v_l\}$，其中，l表示区域特征的数目。此外，为了获得待描述图像的全局性表示，可以通过按位最大操作得到一个池化特征$v_p = \max\{v_l\}_{l=1}^{L}$，将该池化特征作为待描述图像的全局特征。

[0076] S103：将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量。

[0077] 本发明实施例中，文本描述神经网络可以由句子级子网络和词汇级子网络组成，其中，句子级子网络以及词汇级子网络均可以是CNN(Convolutional Neural Networks，卷

积神经网络)。

[0078] 文本描述神经网络是根据训练集预先训练完成的,训练集包括:多个样本图像的区域特征和全局特征,以及每个样本图像对应的样本描述文本及样本描述文本中包含的样本句子数目。

[0079] 为了便于理解,可以参见图3,图3为本发明实施例提供的图像的文本描述方法的一种流程示意图。

[0080] 如图3所示,图片特征提取器提取出的图片特征输入句子级子网络后,能够输出句子指导向量,词汇级子网络根据句子指导向量生成针对待描述图像的段落描述文本。

[0081] S104:将句子指导向量输入文本描述神经网络中的词汇级子网络,得到描述文本。

[0082] 本发明实施例中,句子指导向量能够指导词汇级子网络生成词汇。将句子指导向量输入词汇级子网络,即可得到由多个句子组成的段落文本描述。

[0083] 可见,应用本发明实施例提供的图像的文本描述方法,能够获取待描述图像,提取待描述图像的多个特征区域和一个全局特征,将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络,得到针对每个待生成句子的句子指导向量,将句子指导向量输入文本描述神经网络中的词汇级子网络,得到描述文本。由于采用句子级子网络和词汇级子网络的分层结构,能够捕捉段落中句子之间的连贯性,此外,相比于现有的RNN解码器,降低了计算复杂度。

[0084] 在本发明的一种实施例中,句子级子网络中包含句子嵌入层,门控卷积层和区域感知层,上述步骤S103,具体可以包括以下细化步骤:

[0085] 步骤11:获取当前已生成的每个句子文本;

[0086] 为了便于理解,结合图4进行说明,图4为本发明实施例提供的图像的文本描述方法的另一种流程示意图。

[0087] 如图4所示,本发明实施例中,生成的整个段落描述文本中包含M个句子,表示为:$\hat{S}_1$, $\hat{S}_2$, $\hat{S}_3$.....$\hat{S}_{M-1}$, $\hat{S}_M$,生成段落描述文本的过程中,句子级子 网络能够根据当前已生成的每个句子,确定下一个待生成句子的指导向量,词汇级子网络再根据该指导向量生成下一个句子。

[0088] 其中,句子级子网络包含的句子嵌入层的作用是将当前已生成的句子文本转换为向量的形式。

[0089] 本步骤中,可以确定当前已生成的句子文本。特别的,由于段落中的第一个句子文本之前不存在已生成的句子,因此可以预先设置一个初始句子文本,正如图4中所示的初始句子<S>。

[0090] 步骤12:将每个句子文本输入句子嵌入层,得到每个句子文本的句子嵌入向量;

[0091] 在本发明的一种实施例中,可以将每次生成的句子文本输入句子嵌入层,得到该句子文本对应的句子嵌入向量。

[0092] 参见图4,初始句子<S>输入句子嵌入层后得到第一个句子嵌入向量S_0^e,生成的第一个句子文本$\hat{S}_1$输入句子嵌入层后得到第二个句子嵌入向量S_1^e,以此类推。

[0093] 步骤13:将每个句子嵌入向量均与全局特征的向量拼接,得到多个拼接向量;

[0094] 在本步骤中,可以将当前生成的句子嵌入向量与待描述图像的全局特征的向量拼

接，得到拼接向量。从而，拼接向量既包含了当前已生成句子的特征，又包含了图像的全局特征。

[0095] 举例来讲，如图4所示，待描述图像的全局特征的向量表示为V_p，那么在生成第一个句子文本过程中，V_p与第一个句子嵌入向量S_0^e进行拼接，可以得到第一个拼接向量$I_1=\text{concat}[V_p, S_0^e]$，其中concat表示向量的拼接。

[0096] 步骤14：将每个拼接向量输入门控卷积层，得到针对当前待生成句子的隐藏向量；

[0097] 在本发明实施例中，对于要生成的第i个句子，门控卷积层采用先前已生成的所有句子对应的拼接向量作为输入，生成针对当前待生成的第i个句子的隐藏向量h_i^s。

[0098] 在本发明的一种实施例中，上述步骤14，具体可以包括以下细化步骤：

[0099] 步骤14.a：基于每个拼接向量进行卷积运算，得到针对当前待生成句子的综合语义向量以及门控向量；

[0100] 本发明实施例中，门控卷积层是包含门控线性单元的卷积网络，输入已生成的句子对应的拼接向量，能够输出针对当前待生成句子的综合语义向量h_i^a以及门控向量h_i^b。

[0101] 为了便于说明，以生成第三个句子$\hat{S}_3$的过程为例进行说明。参见图4，将第一个拼接向量I_1，第二个拼接向量I_2以及第三个拼接向量I_3均作为门控卷积层的输入，门控卷积层输出针对待生成的第三个句子的综合语义向量h_3^a及门控向量h_3^b。

[0102] 在本发明的一种实施例中，训练完成的门控卷积层可以基于如下公式生成综合语义向量以及门控向量。

[0103] $$h_i^a = W_a * I_{<i} + b_a$$

[0104] $$h_i^b = W_b * I_{<i} + b_b$$

[0105] 其中，$I_{<i}$表示第i个待生成句子之前已生成的句子对应的拼接向量，W_a，b_a，W_b，b_b均为训练参数。

[0106] 步骤14.b：基于综合语义向量，以及门控向量，进行语义筛选运算，得到针对当前待生成句子的隐藏向量。

[0107] 在本发明的一种实施例中，可以基于如下公式，生成针对当前待生成句子的隐藏向量：

[0108] $$h_i^s = h_i^a \odot \sigma(h_i^b)$$

[0109] 其中，上式所表示的操作为语义筛选运算。具体的，符号⊙表示逐元素乘法运算，σ表示sigmoid函数，即$\sigma(x)=1/(1+e^{-x})$。综合语义向量h_i^a包含了已生成句子的信息，$\sigma(h_i^b)$函数能够选择性的记忆h_i^a中的信息。

[0110] 步骤15：将区域特征的向量和隐藏向量输入区域感知层，得到针对当前待生成句子的句子指导向量。

[0111] 在本发明的一种实施例中，针对当前待生成的第i个句子，可以将步骤15 中得到的隐藏向量h_i^s，以及待描述图像的区域特征的向量共同输入区域感知层，从而得到针对待生成的第i个句子的句子指导向量。

[0112] 其中,上述步骤15可以包括以下细化步骤:

[0113] 步骤15.a:基于区域特征的向量,以及隐藏向量,计算针对每个区域特征的向量的权重;

[0114] 在本发明的一种实施例中,可以基于如下公式,计算针对当前待生成的第 i个句子的各个区域特征的向量的权重:

[0115] $$s_{il}=\frac{\exp((h_i^s)^T W_a v_l)}{\sum_{l=1}^{L}(\exp(h_i^s)^T W_a v_l))}$$

[0116] 其中,v_l表示第l个区域特征的向量,W_a为训练参数,s_{il}表示针对待生成的第i个句子的第l个区域特征的向量的权重。

[0117] 步骤15.b:基于权重,计算针对当前待生成句子的加权区域特征向量;

[0118] 在本发明的一种实施例中,针对每一个待生成句子,可以根据步骤15.a计算得到的权重,计算一个加权区域特征向量。

[0119] 具体的,针对当前待生成的第i个句子,可以基于如下公式,计算加权区域特征向量:

[0120] $$v_i^{att}=\sum_{l=1}^{L} s_{il}\, v_l$$

[0121] 其中,v_i^{att}表示针对当前待生成的第i个句子的加权区域特征向量。

[0122] 步骤15.c:基于加权区域特征向量,以及隐藏向量,计算针对当前待生成句子的句子指导向量。

[0123] 在本发明的一种实施例中,针对当前待生成的第i个句子,基于如下公式计算句子指导向量:

[0124] $$G_i=f(W_{ag}[h_i^s;v_i^{att}]+b_{ag})$$

[0125] 其中,G_i表示针对当前待生成的第i个句子的句子指导向量,f表示激活函数,例如ReLU(Rectified Linear Unit,线性整流函数),W_{ag},b_{ag}均为训练参数。

[0126] 在本发明的一种实施例中,句子级子网络生成针对当前待生成的第i个句子的指导向量G_i后,词汇级子网络可以在指导向量G_i的指导下,为第i个句子生成单词。

[0127] 针对每一个句子中当前要生成单词的预测依赖于该句子中先前已生成的所有单词。特别的,由于每一句话中的第一个词汇之前不存在已生成的词汇,因此可以预先设置一个初始词汇。

[0128] 在本发明的一种实施例中,词汇级子网络中包含词汇嵌入层,其作用是将生成的词汇转换为向量的形式。可以用$w_{i,0}^e$表示第i个句子中第一个词汇,即初始词汇的嵌入向量,用$w_{i,1}^e$表示第i个句子中的第二个词汇,以此类推。

[0129] 在本发明的一种实施例中,在生成第i个句子的第j个词汇时,词汇级子网络以第i个句子的指导向量G_i以及第i个句子先前已生成的词汇的嵌入向量作为输入,即:

[0130] $$h_{i,j}^w=CNN(G_i,w_{i,0},\ldots,w_{i,j-1})$$

[0131] 其中，$h_{i,j}^{w}$表示第i个句子的第j个词汇的隐藏向量。

[0132] 在本发明的一种实施例中，词汇级子网络可以根据隐藏向量，来生成相应的词汇。

[0133] 一种实施例中，可以基于如下公式预测单词的分布：

[0134] $$p_{i,j} = softmax(W_p h_{i,j}^{w})$$

[0135] 其中，W_p表示训练参数，softmax表示一种回归函数，$p_{i,j}$表示第i个句子中第j个单词的预测分布。

[0136] 在本发明的一种实施例中，文本描述神经网络可以采用如下步骤训练得到：

[0137] 步骤21：获取预设的神经网络模型和训练集；

[0138] 步骤22：将样本图像的区域特征和全局特征输入神经网络模型，得到描述文本以及描述文本中包含的句子数目；

[0139] 步骤23：基于得到的描述文本以及句子数目，和训练集中包含的样本描述文本以及样本句子数目，确定损失值；

[0140] 步骤24：根据损失值确定神经网络模型是否收敛；若否，则执行步骤25，若是，则执行步骤26；

[0141] 步骤25：调整神经网络模型中的参数值，并返回步骤22；

[0142] 步骤26：将当前的神经网络模型确定为文本描述神经网络。

[0143] 在本发明的一种实施例中，可以将损失函数定义为句子层的损失和词汇层的损失的加权和，其中，句子层的损失为描述文本段落中句子的个数的损失，词汇层的损失为句子中单词预测分布的损失。

[0144] 具体的，总损失函数$\mathcal{L}$可以表示为：

[0145] $$\mathcal{L} = \lambda_s \sum_{i=1}^{M} \mathcal{L}_s\left(p_i, \mathbb{I}_{i=M}\right) + \lambda_w \sum_{i=1}^{M} \sum_{j=1}^{N_i} \mathcal{L}_w\left(p_{i,j}, \widehat{w}_{i,j}\right)$$

[0146] 其中，M表示句子的总数，N_i表示第i个句子中单词的总数，λ_s表示句子级损失的权重，λ_w表示词汇级损失的权重，$\mathcal{L}_S$表示句子级损失函数，p_i表示第i 个句子的停止分布，符号$\mathbb{I}_{\{\cdot\}}$表示指示函数；$\mathcal{L}_w$表示词汇级损失函数，$\widehat{w}_{i,j}$表示训练样本中第i个句子中第j个单词的真实分布。

[0147] 相应于本发明实施例提供的图像的文本描述方法，本发明实施例还提供了一种图像的文本描述装置，参见图5，图5为本发明实施例提供的图像的文本描述装置的一种结构示意图，装置包括：

[0148] 获取模块501，用于获取待描述图像；

[0149] 提取模块502，用于提取待描述图像的多个区域特征和一个全局特征；

[0150] 第一输入模块503，用于将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量；

[0151] 第二输入模块504，用于将句子指导向量输入文本描述神经网络中的词汇级子网络，得到描述文本；描述文本神经网络是根据训练集训练得到的，训练集包括：多个样本图像的区域特征和全局特征，以及每个样本图像对应的样本描述文本以及样本描述文本中包

含的样本句子数目。

[0152] 在本发明的一种实施例中，句子级子网络中包含句子嵌入层，门控卷积层和区域感知层，第一输入模块503，具体用于：

[0153] 获取当前已生成的每个句子文本；

[0154] 将每个句子文本输入句子嵌入层，得到每个句子文本的句子嵌入向量；

[0155] 将每个句子嵌入向量均与全局特征的向量拼接，得到多个拼接向量；

[0156] 将每个拼接向量输入门控卷积层，得到针对当前待生成句子的隐藏向量；

[0157] 将区域特征的向量和隐藏向量输入区域感知层，得到针对当前待生成句子的句子指导向量。

[0158] 在本发明的一种实施例中，门控卷积层，具体用于：

[0159] 基于每个拼接向量进行卷积运算，得到针对当前待生成句子的综合语义向量以及门控向量；

[0160] 基于综合语义向量，以及门控向量，进行语义筛选运算，得到针对当前待生成句子的隐藏向量。

[0161] 在本发明的一种实施例中，区域感知层，具体用于：

[0162] 基于区域特征的向量，以及隐藏向量，计算针对每个区域特征的向量的权重；

[0163] 基于权重，计算针对当前待生成句子的加权区域特征向量；

[0164] 基于加权区域特征向量，以及隐藏向量，计算针对当前待生成句子的句子指导向量。

[0165] 在本发明的一种实施例中，装置还包括训练模块，训练模块具体用于：

[0166] 获取预设的神经网络模型和训练集；

[0167] 将样本图像的区域特征和全局特征输入神经网络模型，得到描述文本以及描述文本中包含的句子数目；

[0168] 基于得到的描述文本以及句子数目，和训练集中包含的样本描述文本及样本句子数目，确定损失值；

[0169] 根据损失值确定神经网络模型是否收敛；

[0170] 若否，则调整神经网络模型中的参数值，并返回将样本图像的区域特征和全局特征输入神经网络模型，得到描述文本以及描述文本中包含的句子数目的步骤；

[0171] 若是，则将当前的神经网络模型确定为文本描述神经网络。

[0172] 本发明实施例提供的图像的文本描述装置，能够获取待描述图像，提取待描述图像的多个区域特征和一个全局特征；将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量；将句子指导向量输入文本描述神经网络中的词汇子网络，得到描述文本；由于采用句子级子网络和词汇级子网络的分层结构，能够捕捉段落中句子之间的连贯性，提高了生成的文本段落中句子之间的连贯性，此外，相较于现有的未进行分级训练的方案，降低了训练过程的计算复杂度。

[0173] 本发明实施例还提供了一种电子设备，如图6所示，包括处理器601、通信接口602、存储器603和通信总线604，其中，处理器601，通信接口602，存储器603通过通信总线604完成相互间的通信，

[0174] 存储器603，用于存放计算机程序；

[0175] 处理器601,用于执行存储器603上所存放的程序时,实现如下步骤:

[0176] 获取待描述图像;

[0177] 提取待描述图像的多个区域特征和一个全局特征;

[0178] 将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络,得到针对每个待生成句子的句子指导向量;

[0179] 将句子指导向量输入文本描述神经网络中的词汇级子网络,得到描述文本。

[0180] 上述电子设备提到的通信总线可以是外设部件互连标准(Peripheral Component Interconnect,PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture,EISA)总线等。该通信总线可以分为地址总线、数据总线、控制总线等。为便于表示,图6中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0181] 通信接口用于上述电子设备与其他设备之间的通信。

[0182] 存储器可以包括随机存取存储器(Random Access Memory,RAM),也可以包括非易失性存储器(Non-Volatile Memory,NVM),例如至少一个磁盘存储器。可选的,存储器还可以是至少一个位于远离前述处理器的存储装置。

[0183] 上述的处理器可以是通用处理器,包括中央处理器(Central Processing Unit,CPU)、网络处理器(Network Processor,NP)等;还可以是数字信号处理器 (Digital Signal Processing,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现场可编程门阵列(Field-Programmable Gate Array,FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

[0184] 基于相同的发明构思,根据上述图像的文本描述方法实施例,在本发明提供的又一实施例中,还提供了一种计算机可读存储介质,该计算机可读存储介质内存储有计算机程序,计算机程序被处理器执行时实现上述图1,3和4所示的任一图像的文本描述方法步骤。

[0185] 需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0186] 本说明书中的各个实施例均采用相关的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于装置、电子设备以及存储介质实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0187] 以上仅为本发明的较佳实施例而已,并非用于限定本发明的保护范围。凡在本发明的精神和原则之内所作的任何修改、等同替换、改进等,均包含在本发明的保护范围内。

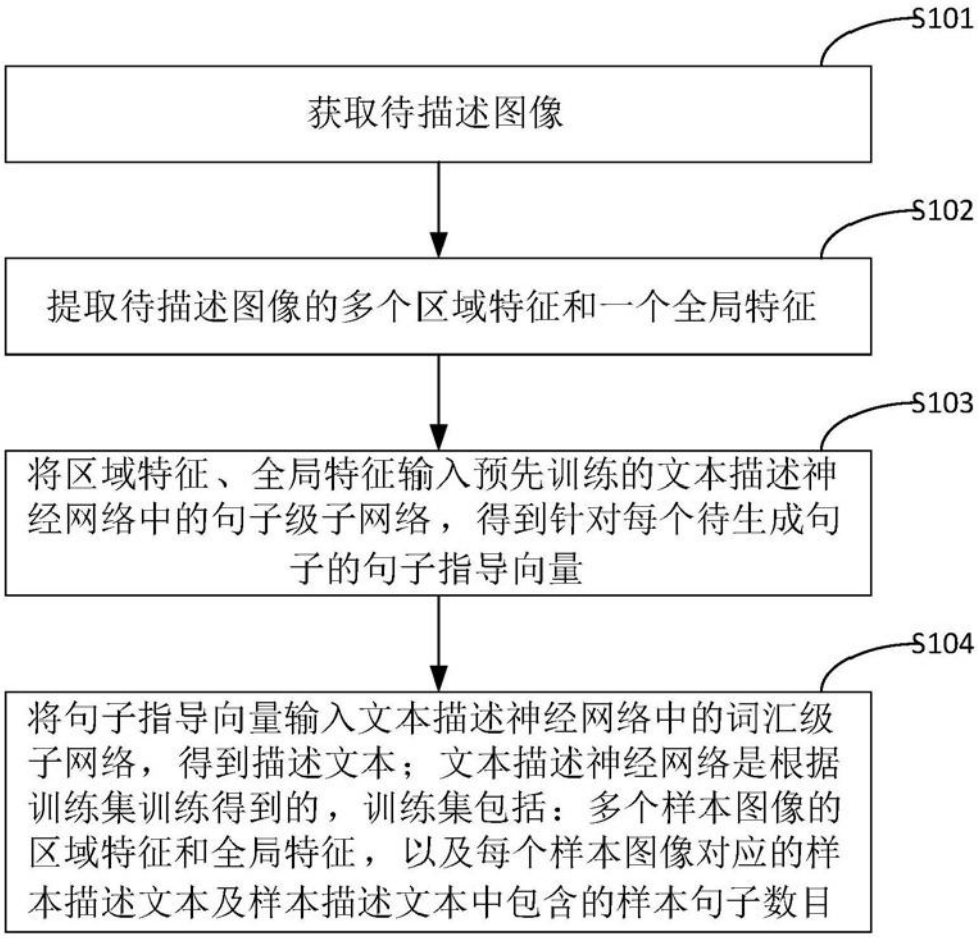

S101
获取待描述图像
S102
提取待描述图像的多个区域特征和一个全局特征
S103
将区域特征、全局特征输入预先训练的文本描述神经网络中的句子级子网络，得到针对每个待生成句子的句子指导向量
S104
将句子指导向量输入文本描述神经网络中的词汇级子网络，得到描述文本；文本描述神经网络是根据训练集训练得到的，训练集包括：多个样本图像的区域特征和全局特征，以及每个样本图像对应的样本描述文本及样本描述文本中包含的样本句子数目

图1

图2

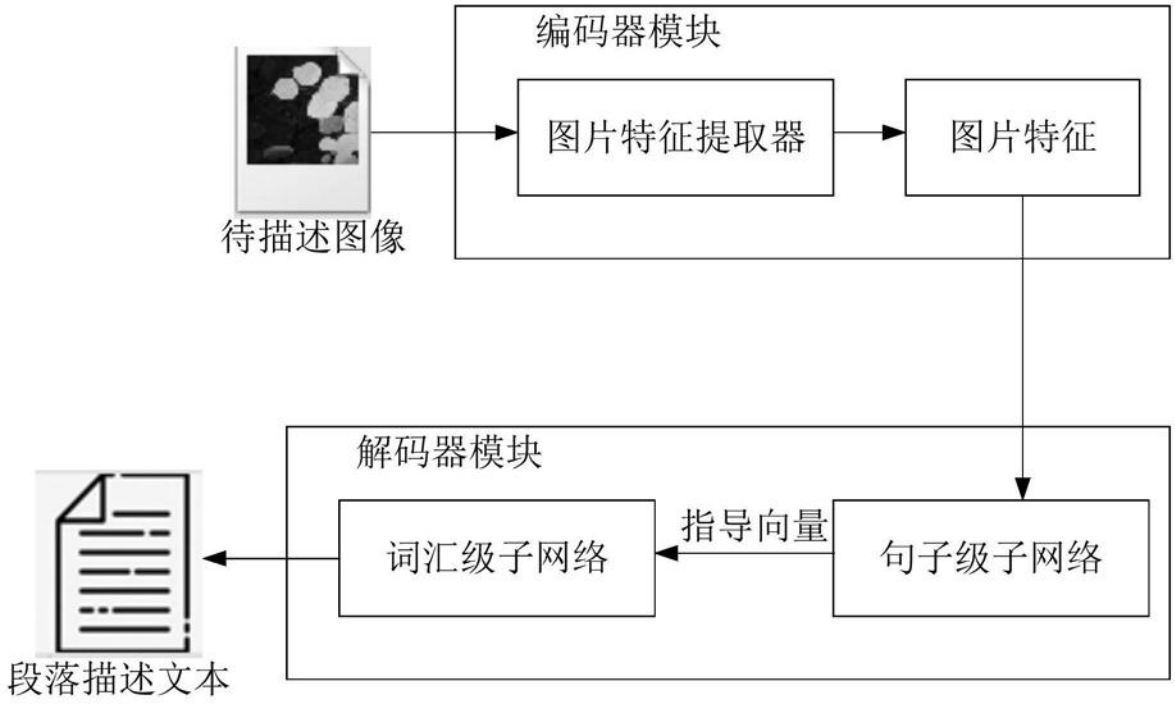

编码器模块
图片特征提取器
图片特征
待描述图像
解码器模块
词汇级子网络
指导向量
句子级子网络
段落描述文本

图3

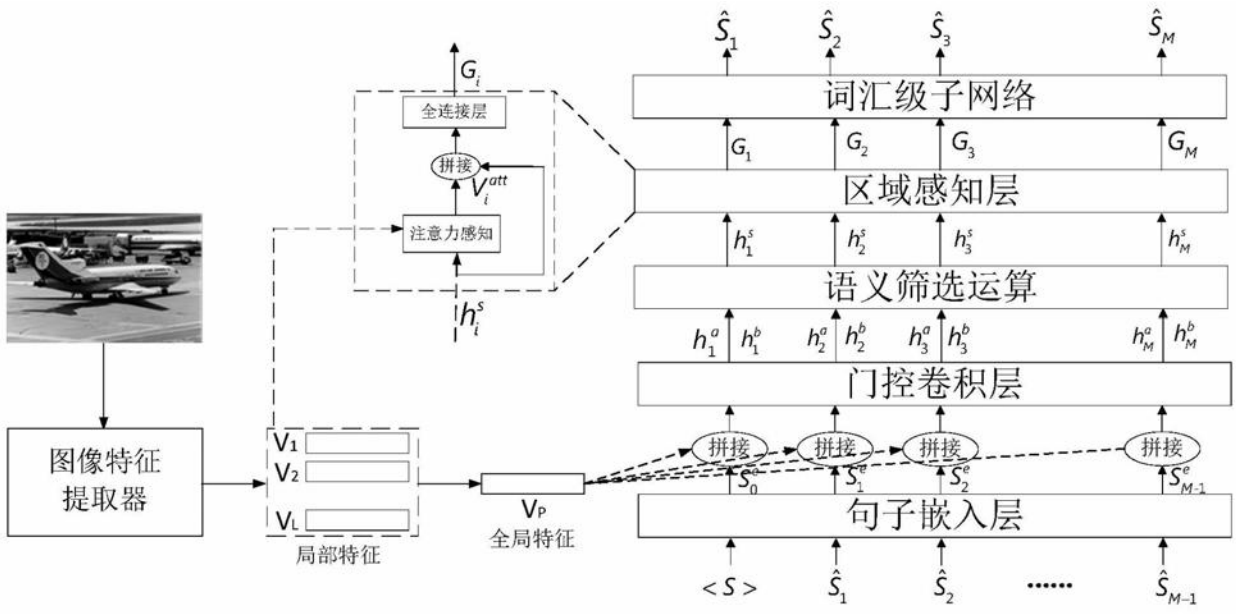

全连接层
拼接
注意力感知
图像特征提取器
局部特征
全局特征
词汇级子网络
区域感知层
语义筛选运算
门控卷积层
句子嵌入层
拼接

图4

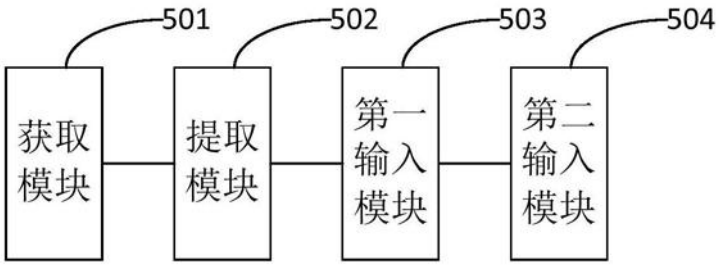

501
502
503
504
获取模块
提取模块
第一输入模块
第二输入模块

图5

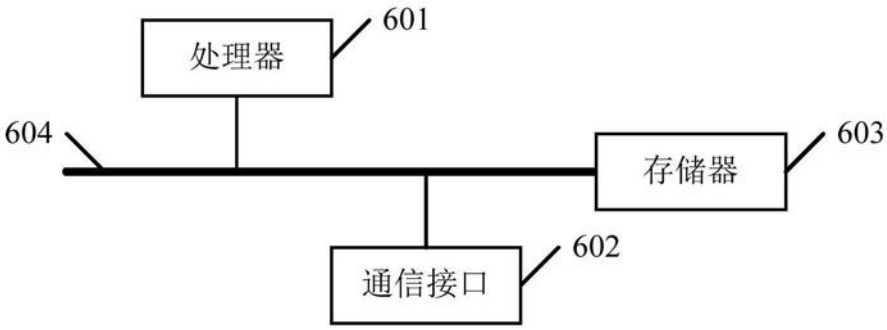

601
处理器
604
603
存储器
602
通信接口

图6