(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 114266905 A
(43)申请公布日 2022.04.01

(21)申请号 202210028340.X
(22)申请日 2022.01.11
(71)申请人 重庆师范大学
地址 401331 重庆市沙坪坝区大学城中路37号
(72)发明人 翟浩 陈立志 方小龙 潘龙越 杨有
(74)专利代理机构 北京和联顺知识产权代理有限公司 11621
代理人 白京萍
(51)Int.Cl.
G06V 10/46(2022.01)
G06V 10/44(2022.01)
G06V 10/764(2022.01)
G06V 10/82(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

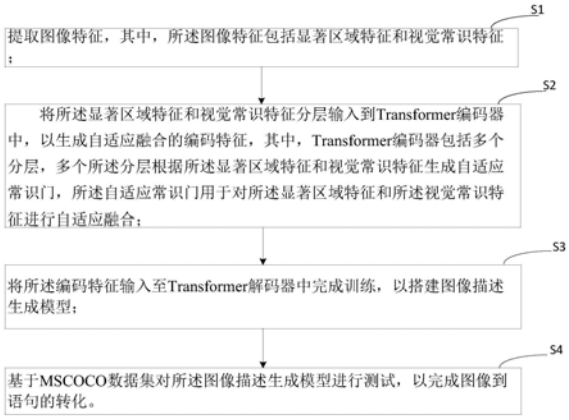
权利要求书3页 说明书12页 附图3页

(54)发明名称

基于Transformer结构的图像描述生成模型方法、装置和计算机设备

(57)摘要

本申请涉及计算机视觉和自然语言处理技术领域，公开了一种基于Transformer结构的图像描述生成模型方法、装置和计算机设备，本申请使用Faster R-CNN模型提取图像显著区域特征，使用VC R-CNN模型提取视觉常识特征，通过将显著区域特征和视觉常识特征分层输入到Transformer编码器中，并在每一分层中设计使用了自适应常识门，从而增强了图像描述生成模型对视觉常识信息的提取能力，同时进一步融合了图像的显著区域信息和视觉常识信息，生成更加符合语境的描述语句，从而减少生成语句中的内容缺失，提高描述语句的准确性。



CN 114266905 A

1.一种基于Transformer结构的图像描述生成模型方法,其特征在于,包括:

提取图像特征,其中,所述图像特征包括显著区域特征和视觉常识特征;

将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中,以生成自适应融合的编码特征,其中,Transformer编码器包括多个分层,多个所述分层根据所述显著区域特征和视觉常识特征生成自适应常识门,所述自适应常识门用于对所述显著区域特征和所述视觉常识特征进行自适应融合;

将所述编码特征输入至Transformer解码器中完成训练,以搭建图像描述生成模型;

基于MSCOCO数据集对所述图像描述生成模型进行测试,以完成图像到语句的转化。

2.根据权利要求1所述的基于Transformer结构的图像描述生成模型方法,其特征在于,所述提取图像特征的步骤,包括:

基于Faster R-CNN构建图像的区域建议网络;

将所述区域建议网络引入深度神经网络模型中,得到基于深度神经网络的组合图像特征,将所述组合图像特征作为显著区域特征;

基于VCR-CNN提取图像边界框的坐标,其中,坐标包括多个;

将多个所述坐标输入卷积神经网络模型中进行训练,训练完成后得到视觉常识特征。

3.根据权利要求2所述的基于Transformer结构的图像描述生成模型方法,其特征在于,所述将所述区域建议网络引入深度神经网络模型中,得到基于深度神经网络的组合图像特征的步骤,包括:

基于所述区域建议网络获取多个不同批次的多个第一图像;

对每一个批次的每一个所述第一图像进行短边缩放,得到每一个批次的短边缩放的多个第二图像;

将每一个批次的多个所述第二图像传入卷积神经网络层中以对多个所述第二图像进行卷积和池化,以生成每一个批次的多个第二图像的组合图像特征。

4.根据权利要求1所述的基于Transformer结构的图像描述生成模型方法,其特征在于,所述将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中,以生成自适应融合的编码特征的步骤,包括:

对所述显著区域特征和所述视觉常识特征进行拼接,得到拼接融合特征;

根据所述拼接融合特征对图像模态间和模态内的常识性关系进行建模,得到ACG融合模型;

将所述拼接融合特征输入到所述ACG融合模型中进行训练,得到ACG输出特征;

将所述ACG输出特征分层输入到自注意力块中进行融合,得到多个层次的融合编码向量,其中,所述自注意力块包括多个,多个所述自注意力块进行模态内和跨模态的分层交互;

对所述融合编码向量进行残差和归一化处理,得到自适应融合的编码特征。

5.根据权利要求4所述的基于Transformer结构的图像描述生成模型方法,其特征在于,所述对所述显著区域特征和所述视觉常识特征进行拼接,得到拼接融合特征的步骤,包括:

基于所述视觉常识特征依次获取每一个所述视觉常识特征对应的视觉特征向量;

基于所述显著区域特征依次获取每一个所述显著区域特征对应的显著区域向量;

根据所述视觉特征向量与所述显著区域向量对所述显著区域特征和所述视觉常识特征进行拼接，其中，拼接公式为：

$vc_i=[v_i,c_i]$；

$VC=\{vc_1,vc_2,\cdots,vc_N\}$；

其中，所述$v_i\in R^d$，$c_i\in R^d$，$vc_i\in R^{2d}$，$v_i\in R^d$表示d维的第i个视觉特征向量，$c_i\in R^d$表示d维的第i个显著区域向量，$vc_i\in R^{2d}$表示2d维的显著区域特征和视觉常识特征的拼接向量；所述vc_i表示第i个显著区域特征和第i个视觉常识特征拼接；VC表示拼接融合特征。

6.根据权利要求4所述的基于Transformer结构的图像描述生成模型方法，其特征在于，根据所述拼接融合特征对图像模态间和模态内的常识性关系进行建模，得到ACG融合模型的步骤，包括：

获取所述拼接融合特征中显著区域特征的第一线性表示；

获取所述拼接融合特征中视觉常识特征的第二线性表示；

根据所述第一线性表示与所述第二线性表示，计算拼接融合特征的线性表示施加影响，其中，计算公式为：

$f_{vc}=sigmoid(g_v+g_c)*g_c$；

其中，所述f_{vc}表示拼接融合特征的线性表示施加影响；g_v表示第一线性表示；g_c表示第二线性表示；

根据所述线性表示施加影响对所述拼接融合特征的模态间和模态内的常识性关系进行建模，得到ACG融合模型，其中，建模过程为：

$V_{acg}=\tanh(W_f f_{vc}+b_f)+V$；

其中，V_{acg}表示模态间和模态内的常识性关系，f_{vc}表示拼接融合特征的线性表示施加影响，W_f表示需要被学习的权重，b_f表示偏置项，V表示显著区域特征，$V=\{v_1,v_2,\cdots,v_N\}$。

7.根据权利要求1所述的基于Transformer结构的图像描述生成模型方法，其特征在于，将所述编码特征输入至Transformer解码器中完成训练，以搭建图像描述生成模型的步骤，包括：

向Transformer解码器中的掩码自注意块输入标签信息，并将所述掩码自注意块作为第一子层，得到第一子层的第一特征信息；

将所述第一特征信息与所述编码特征作为查询向量输入到Transformer解码器中的交叉注意力块中，并将所述交叉注意力块作为第二子层，得到第二子层的第二特征信息；

将所述第二特征信息输入到位置前馈网络进行非线性变换训练；

返回到所述向Transformer解码器中的掩码自注意块输入标签信息的步骤，并对返回次数进行计数，得到返回总数；

判断所述返回总数是否超过预设次数；

若所述返回总数超过预设次数，判定所述第二特征信息训练完成，搭建图像描述生成模型。

8.一种基于Transformer结构的图像描述生成模型装置，其特征在于，包括：

提取模块，用于提取图像特征，其中，所述图像特征包括显著区域特征和视觉常识特征；

分层输入模块，用于将所述显著区域特征和视觉常识特征分层输入到Transformer编

码器中,以生成自适应融合的编码特征,其中,Transformer编码器包括多个分层,多个所述分层根据所述显著区域特征和视觉常识特征生成自适应常识门,所述自适应常识门用于对所述显著区域特征和所述视觉常识特征进行自适应融合;

训练模块,用于将所述编码特征输入至Transformer解码器中完成训练,以搭建图像描述生成模型;

测试模块,用于基于MSCOCO数据集对所述图像描述生成模型进行测试,以完成图像到语句的转化。

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中任一项所述基于Transformer结构的图像描述生成模型方法的步骤。

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中任一项所述基于Transformer结构的图像描述生成模型方法的步骤。

基于Transformer结构的图像描述生成模型方法、装置和计算机设备

技术领域

[0001] 本申请涉及计算机视觉和自然语言处理技术领域，特别涉及一种基于Transformer结构的图像描述生成模型方法、装置和计算机设备。

背景技术

[0002] 图像描述生成是一个融合了计算机视觉和自然语言处理的跨模态任务，它是图像处理的高级层次。从传统的基于检索、模板的方法到后来基于编码器-解码器的深度学习方法，使图像描述生成任务从只能生成单一形式的语句发展到现在可以生成精度更高、描述更加准确的语句。在常规的编解码框架中，使用了卷积神经网络(Convolutional Neural Network，CNN)作为编码器将图像编码为相应特征，使用长短期记忆(Long Short-Term Memory，LSTM)网络作为解码器将图像特征解码为对应描述句子。现有技术中首先提出了一种完全依赖于注意力机制的Transformer结构，可以对图像特征进行关系建模，解决了循环神经网络(Recurrent Neural Networks，RNN)存在的时间依赖问题。随后，基于Transformer结构的各种变体在图像描述模型中应运而生。2018年，又首次采用Faster R-CNN(Faster Region-based Convolutional Neural Network)作为编码器，提取图像的显著区域特征并应用在图像描述生成模型中，在Transformer结构之上，通过几何注意合并对象事物之间的空间关系信息，证明了模型空间意识的重要性。以上的图像描述生成模型虽然能产生描述图像语义内容的句子，但无法确切描述图像事物的因果关系，存在潜在的视觉注意不合理问题。且图像描述生成模型在使用多个特征进行融合处理时，会出现特征信息利用不充分且融合方式单一的问题，存在一定的局限性，例如，整体融合方式是单一拼接，没有重点融合其中的关键信息，这导致生成语句存在重要内容缺失问题。

发明内容

[0003] 本申请的主要目的为提供一种基于Transformer结构的图像描述生成模型方法，旨在解决现有技术中图像描述生成模型在使用多个特征进行融合处理时，会出现特征信息利用不充分且融合方式单一，导致生成语句存在重要内容缺失的技术问题。

[0004] 本申请提出一种基于Transformer结构的图像描述生成模型方法，包括：

[0005] 提取图像特征，其中，所述图像特征包括显著区域特征和视觉常识特征；

[0006] 将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中，以生成自适应融合的编码特征，其中，Transformer编码器包括多个分层，多个所述分层根据所述显著区域特征和视觉常识特征生成自适应常识门，所述自适应常识门用于对所述显著区域特征和所述视觉常识特征进行自适应融合；

[0007] 将所述编码特征输入至Transformer解码器中完成训练，以搭建图像描述生成模型；

[0008] 基于MSCOCO数据集对所述图像描述生成模型进行测试，以完成图像到语句的转

化。

[0009] 作为优选,所述提取图像特征的步骤,包括:

[0010] 基于Faster R-CNN构建图像的区域建议网络;

[0011] 将所述区域建议网络引入深度神经网络模型中,得到基于深度神经网络的组合图像特征,将所述组合图像特征作为显著区域特征;

[0012] 基于VCR-CNN提取图像边界框的坐标,其中,坐标包括多个;

[0013] 将多个所述坐标输入卷积神经网络模型中进行训练,训练完成后得到视觉常识特征。

[0014] 作为优选,所述将所述区域建议网络引入深度神经网络模型中,得到基于深度神经网络的组合图像特征的步骤,包括:

[0015] 基于所述区域建议网络获取多个不同批次的多个第一图像;

[0016] 对每一个批次的每一个所述第一图像进行短边缩放,得到每一个批次的短边缩放的多个第二图像;

[0017] 将每一个批次的多个所述第二图像传入卷积神经网络层中以对多个所述第二图像进行卷积和池化,以生成每一个批次的多个第二图像的组合图像特征。

[0018] 作为优选,所述将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中,以生成自适应融合的编码特征的步骤,包括:

[0019] 对所述显著区域特征和所述视觉常识特征进行拼接,得到拼接融合特征;

[0020] 根据所述拼接融合特征对图像模态间和模态内的常识性关系进行建模,得到ACG融合模型;

[0021] 将所述拼接融合特征输入到所述ACG融合模型中进行训练,得到ACG输出特征;

[0022] 将所述ACG输出特征分层输入到自注意力块中进行融合,得到多个层次的融合编码向量,其中,所述自注意力块包括多个,多个所述自注意力块进行模态内和跨模态的分层交互;

[0023] 对所述融合编码向量进行残差和归一化处理,得到自适应融合的编码特征。

[0024] 作为优选,所述对所述显著区域特征和所述视觉常识特征进行拼接,得到拼接融合特征的步骤,包括:

[0025] 基于所述视觉常识特征依次获取每一个所述视觉常识特征对应的视觉特征向量;

[0026] 基于所述显著区域特征依次获取每一个所述显著区域特征对应的显著区域向量;

[0027] 根据所述视觉特征向量与所述显著区域向量对所述显著区域特征和所述视觉常识特征进行拼接,其中,拼接公式为:

[0028] $vc_i=[v_i,c_i]$;

[0029] $VC=\{vc_1,vc_2,\cdots,vc_N\}$;

[0030] 其中,所述$v_i\in R^d$,$c_i\in R^d$,$vc_i\in R^{2d}$,$v_i\in R^d$表示d维的第i个视觉特征向量,$c_i\in R^d$表示d维的第i个显著区域向量,$vc_i\in R^{2d}$表示2d维的显著区域特征和视觉常识特征的拼接向量;所述vc_i表示第i个显著区域特征和第i个视觉常识特征拼接;VC表示拼接融合特征。

[0031] 作为优选,根据所述拼接融合特征对图像模态间和模态内的常识性关系进行建模,得到ACG融合模型的步骤,包括:

[0032] 获取所述拼接融合特征中显著区域特征的第一线性表示;

[0033] 获取所述拼接融合特征中视觉常识特征的第二线性表示；

[0034] 根据所述第一线性表示与所述第二线性表示，计算拼接融合特征的线性表示施加影响，其中，计算公式为：

[0035] $f_{vc}=\mathrm{sigmoid}(g_v+g_c)*g_c$；

[0036] 其中，所述f_{vc}表示拼接融合特征的线性表示施加影响；g_v表示第一线性表示；g_c表示第二线性表示；

[0037] 根据所述线性表示施加影响对所述拼接融合特征的模态间和模态内的常识性关系进行建模，得到ACG融合模型，其中，建模过程为：

[0038] $V_{acg}=\tanh(W_f f_{vc}+b_f)+V$；

[0039] 其中，V_{acg}表示模态间和模态内的常识性关系，f_{vc}表示拼接融合特征的线性表示施加影响，W_f表示需要被学习的权重，b_f表示偏置项，V表示显著区域特征，$V=\{v_1,v_2,\cdots,v_N\}$。

[0040] 作为优选，将所述编码特征输入至Transformer解码器中完成训练，以搭建图像描述生成模型的步骤，包括：

[0041] 向Transformer解码器中的掩码自注意块输入标签信息，并将所述掩码自注意块作为第一子层，得到第一子层的第一特征信息；

[0042] 将所述第一特征信息与所述编码特征作为查询向量输入到Transformer解码器中的交叉注意力块中，并将所述交叉注意力块作为第二子层，得到第二子层的第二特征信息；

[0043] 将所述第二特征信息输入到位置前馈网络进行非线性变换训练；

[0044] 返回到所述向Transformer解码器中的掩码自注意块输入标签信息的步骤，并对返回次数进行计数，得到返回总数；

[0045] 判断所述返回总数是否超过预设次数；

[0046] 若所述返回总数超过预设次数，判定所述第二特征信息训练完成，搭建图像描述生成模型。

[0047] 本申请还提供一种基于Transformer结构的图像描述生成模型装置，包括：

[0048] 提取模块，用于提取图像特征，其中，所述图像特征包括显著区域特征和视觉常识特征；

[0049] 分层输入模块，用于将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中，以生成自适应融合的编码特征，其中，Transformer编码器包括多个分层，多个所述分层根据所述显著区域特征和视觉常识特征生成自适应常识门，所述自适应常识门用于对所述显著区域特征和所述视觉常识特征进行自适应融合；

[0050] 训练模块，用于将所述编码特征输入至Transformer解码器中完成训练，以搭建图像描述生成模型；

[0051] 测试模块，用于基于MSCOCO数据集对所述图像描述生成模型进行测试，以完成图像到语句的转化。

[0052] 本申请还提供了一种计算机设备，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上述基于Transformer结构的图像描述生成模型方法的步骤。

[0053] 本申请还提供了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上述基于Transformer结构的图像描述生成模型方法的步骤。

[0054] 本申请的有益效果为:本申请使用Faster R-CNN模型提取图像显著区域特征,使用VC R-CNN模型提取视觉常识特征,通过将显著区域特征和视觉常识特征分层输入到Transformer编码器中,并在每一分层中设计使用了自适应常识门,从而增强了图像描述生成模型对视觉常识信息的提取能力,同时进一步融合了图像的显著区域信息和视觉常识信息,生成更加符合语境的描述语句,从而减少生成语句中的内容缺失,提高描述语句的准确性。

附图说明

[0055] 图1为本申请一实施例的基于Transformer结构的图像描述生成模型方法流程示意图。

[0056] 图2为本申请一实施例的基于Transformer结构的图像描述生成模型装置结构示意图。

[0057] 图3为本申请一实施例的计算机设备内部结构示意图。

[0058] 本申请目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

具体实施方式

[0059] 应当理解,此处所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。

[0060] 如图1-图3所示,本申请提出一种基于Transformer结构的图像描述生成模型方法,包括:

[0061] S1、提取图像特征,其中,所述图像特征包括显著区域特征和视觉常识特征;

[0062] S2、将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中,以生成自适应融合的编码特征,其中,Transformer编码器包括多个分层,多个所述分层根据所述显著区域特征和视觉常识特征生成自适应常识门,所述自适应常识门用于对所述显著区域特征和所述视觉常识特征进行自适应融合;

[0063] S3、将所述编码特征输入至Transformer解码器中完成训练,以搭建图像描述生成模型;

[0064] S4、基于MSCOCO数据集对所述图像描述生成模型进行测试,以完成图像到语句的转化。

[0065] 如上述步骤S1-S4所述,通过对图像特征进行提取,再将图像特征输入Transformer编码器中,Transformer编码器的每一个分层中设计了自适应常识门,该自适应常识门可对图像特征进行融合操作,得到自适应融合的编码特征,将编码特征输入到Transformer解码器中完成训练,从而搭建图像描述生成模型,最后基于MSCOCO数据集对所述图像描述生成模型进行测试,以完成图像到语句的转化;本实施例通过自适应常识门机制,能够提高对图像深层关系的感知和表征性,将图像特征分层输入到Transformer编码器中,从而使得Transformer编码器更好的从图像特征中感知图像的因果关系,这样不仅可以降低图像描述生成模型的收敛速度,而且能够充分融合图像特征和编码特征,生成更加符合语境的描述语句,从而减少生成语句中的内容缺失,提高描述语句的准确性。

[0066] 在一个实施例中,所述提取图像特征的步骤S1,包括:

[0067] S11、基于Faster R-CNN构建图像的区域建议网络;

[0068] S12、将所述区域建议网络引入深度神经网络模型中，得到基于深度神经网络的组合图像特征，将所述组合图像特征作为显著区域特征；

[0069] S13、基于VCR-CNN提取图像边界框的坐标，其中，坐标包括多个；

[0070] S14、将多个所述坐标输入卷积神经网络模型中进行训练，训练完成后得到视觉常识特征。

[0071] 如上述步骤S11-S14所述，可通过目标检测网络，基于Faster R-CNN提取图像的显著区域特征以及基于VC R-CNN提取图像的视觉常识特征，从而将图像中的重要特征依次提取出来，这样能够较大程度上对图像的特征进行有效提取，便于后续基于显著区域特征与视觉常识特征进行建模训练。

[0072] 在一个实施例中，所述将所述区域建议网络引入深度神经网络模型中，得到基于深度神经网络的组合图像特征的步骤S12，包括：

[0073] S121、基于所述区域建议网络获取多个不同批次的多个第一图像；

[0074] S122、对每一个批次的每一个所述第一图像进行短边缩放，得到每一个批次的短边缩放的多个第二图像；

[0075] S123、将每一个批次的多个所述第二图像传入卷积神经网络层中以对多个所述第二图像进行卷积和池化，以生成每一个批次的多个第二图像的组合图像特征。

[0076] 如上述步骤S121-S123所述，首先按照配置从区域建议网络中获取一个epoch(epoch:批次数，1个epoch等于使用训练集中的全部样本训练一次；一个epoch＝所有训练样本的一个正向传递和一个反向传递)的多个第一图像，并将短第一图像的短边缩放至600像素，得到每一个批次的短边缩放的多个第二图像，对多个第二图像进行重新排序，再将第二图像传入卷积神经网络层，经过卷积和池化，通过加空白使得每次卷积后大小不变，池化后减半，从而生成每一个批次的多个第二图像的组合图像特征。除此之外，基于VCR-CNN提取视觉常识特征时，可通过干扰因子字典(Confounder Dictionary，CD)存储常识来提取，它的提取方式和Faster R-CNN的提取方式相比，去除了RPN网络，不再训练网络建议边界框，而是直接将训练集中真实词的边界框坐标输入到其中，直接提取区域特征。在训练完成后的特征提取阶段，只要给定图片和边界框坐标，便可以获得对应的视觉常识特征$C=\{c_1, c_2, \ldots, c_N\}$，因此VCR-CNN作为一种改进的视觉区域编码器，可使用因果关系干预该区域的上下文对象，这样能够更加快速的将视觉常识特征提取出来。

[0077] 在一个实施例中，所述将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中，以生成自适应融合的编码特征的步骤S2，包括：

[0078] S21、对所述显著区域特征和所述视觉常识特征进行拼接，得到拼接融合特征；

[0079] S22、根据所述拼接融合特征对图像模态间和模态内的常识性关系进行建模，得到ACG融合模型；

[0080] S23、将所述拼接融合特征输入到所述ACG融合模型中进行训练，得到ACG输出特征；

[0081] S24、将所述ACG输出特征分层输入到自注意力块中进行融合，得到多个层次的融合编码向量，其中，所述自注意力块包括多个，多个所述自注意力块进行模态内和跨模态的分层交互；

[0082] S25、对所述融合编码向量进行残差和归一化处理，得到自适应融合的编码特征。

[0083] 如上述步骤S21-S25所述，通过对显著区域特征和视觉常识特征进行拼接，再基于拼接融合特征对图像模态间和模态内的常识性关系进行建模，得到ACG融合模型，这样能够在ACG融合模型中建立信道特征依赖关系，从而更加有效的指导多通道信息之间的交互，再将拼接融合特征作为训练样本输入到ACG融合模型中进行训练，得到ACG输出特征，将ACG输出特征以分层的方式输入到自注意力块中进行融合，其中，在Transformer编码器中，可利用自注意块进行模态内和跨模态的交互，具体的，可利用多个自注意力块允许ACG输出特征在不同的子空间内学习到相关的信息，且与仅利用一层的多模态交互的注意力模块相比，使用多个自注意力块串联，可以获得不同层次的ACG输出特征，提高了多模态交互的能力，所运用的公式为：$f_{mh\text{-}att}(Q,K,V)=Concat(head_1,head_2,...,head_h)$；$head_i=f_{dot\text{-}att}(Q_i,K_i,V_i)$；$f_{dot\text{-}att}(Q_i,K_i,V_i)=\mathrm{softmax}(\frac{Q_iK_i^{T}}{\sqrt{d}})V_i$；其中：Q、K、V是三个独立线性映射；$f_{mh\text{-}att}$是由n＝8个$f_{dot\text{-}att}$函数组成的多头注意力函数。在Transformer编码器中，自注意模块并行多次重复计算，其中的每一个自注意模块都称为注意头，Attention模块将其Query，Key和value参数进行N次拆分，并将每次拆分分别通过单独的传递，然后将所有这些相似的注意力计算合并在一起以产生最终的注意力得分，这样可以为每个融合编码向量编辑多个关系和细微差别，即对于每对(Q_i,K_i,V_i)，Transformer编码器都能将输入的ACG输出特征映射到不同的子空间，通过多个自注意力块串联从而增加不同子空间的组合关系，进一步增强Transformer编码器的视觉表达能力，更优的，将所述ACG输出特征分层输入到自注意力块中进行融合的步骤，还包括：使用自注意力操作来获得V_{acg}的置换不变编码，该操作的公式为：

[0084] $S(V_{acg})=f_{mh\text{-}att}(V_{acg}W_q,V_{acg}W_k,V_{acg}W_v)$，其中：$W_q\in\mathbf{R}^{d_{model}\times d_k}$、$W_k\in\mathbf{R}^{d_{model}\times d_k}$、$W_v\in\mathbf{R}^{d_{model}\times d_v}$；$V_{acg}$显著区域特征和视觉常识特征组合，$W_q\in\mathbf{R}^{d_{model}\times d_q}$表示$d_{model}\times d_q$维学习的权重，$W_k\in\mathbf{R}^{d_{model}\times d_k}$表示$d_{model}\times d_k$维学习的权重，$W_v\in\mathbf{R}^{d_{model}\times d_v}$表示$d_{model}\times d_k$维学习的权重；自注意力块的输出是一组新的融合编码向量$S(V_{acg})$，其维数(独立的时空坐标的数目)与V_{acg}相同，ACG输出特征和自注意力块组成其中一个子层，本实施例中，在Transformer编码端设置了3个不同的子层，从而得到多个层次的融合编码向量，为了防止模型出现梯度消失问题，采用残差和归一化的方式，对所述融合编码向量进行残差和归一化处理，特征V_{acg}经过一个自注意力块后的输出记为得到自适应融合的编码特征$\tilde{V}_{acg}$，其中，归一化公式为：$\tilde{V}_{acg}=\mathrm{LayerNorm}(V_{acg}+S(V_{acg}))$；其中，LayerNorm为归一化层，$S(V_{acg})$表示输入$V_i$，该$V_i$经过$f_{mh\text{-}att}(Q,K,V)=Concat(head_1,head_2,...,head_h)$；$head_i=f_{dot\text{-}att}(Q_i,K_i,V_i)$；$f_{dot\text{-}att}(Q_i,K_i,V_i)=\mathrm{softmax}(\frac{Q_iK_i^{T}}{\sqrt{d}})V_i$的输出；这样能够保留重要的图像常识信息，建立显著区域特征和视觉常识特征之间的依赖关系。

[0085] 在一个实施例中，所述对所述显著区域特征和所述视觉常识特征进行拼接，得到拼接融合特征的步骤S21，包括：

[0086] S211、基于所述视觉常识特征依次获取每一个所述视觉常识特征对应的视觉特征向量；

[0087] S212、基于所述显著区域特征依次获取每一个所述显著区域特征对应的显著区域向量；

[0088] S213、根据所述视觉特征向量与所述显著区域向量对所述显著区域特征和所述视觉常识特征进行拼接，其中，拼接公式为：

[0089] $vc_i=[v_i,c_i]$；

[0090] $VC=\{vc_1,vc_2,\cdots,vc_N\}$；

[0091] 其中，所述$v_i\in R^d$，$c_i\in R^d$，$vc_i\in R^{2d}$，$v_i\in R^d$表示d维的第i个视觉特征向量，$c_i\in R^d$表示d维的第i个显著区域向量，$vc_i\in R^{2d}$表示2d维的显著区域特征和视觉常识特征的拼接向量；所述vc_i表示第i个显著区域特征和第i个视觉常识特征拼接；VC表示拼接融合特征。

[0092] 如上述步骤S211-S213所述，可基于$vc_i=[v_i,c_i]$以及$C=\{vc_1,vc_2,\cdots,vc_N\}$从而对显著区域特征和所述视觉常识特征进行拼接，这样能够建立显著区域特征和视觉常识特征之间的特征关系，得到拼接融合特征，便于从拼接融合特征中感知图像的因果关系。

[0093] 在一个实施例中，根据所述拼接融合特征对图像模态间和模态内的常识性关系进行建模，得到ACG融合模型的步骤S22，包括：

[0094] S221、获取所述拼接融合特征中显著区域特征的第一线性表示；

[0095] S222、获取所述拼接融合特征中视觉常识特征的第二线性表示；

[0096] S223、根据所述第一线性表示与所述第二线性表示，计算拼接融合特征的线性表示施加影响，其中，计算公式为：

[0097] $f_{vc}=\mathrm{sigmoid}(g_v+g_c)*g_c$；

[0098] 其中，所述f_{vc}表示拼接融合特征的线性表示施加影响；g_v表示第一线性表示；g_c表示第二线性表示；

[0099] S224、根据所述线性表示施加影响对所述拼接融合特征的模态间和模态内的常识性关系进行建模，得到ACG融合模型，其中，建模过程为：

[0100] $V_{acg}=\tanh(W_f f_{vc}+b_f)+V$；

[0101] 其中，V_{acg}表示模态间和模态内的常识性关系，f_{vc}表示拼接融合特征的线性表示施加影响，W_f表示需要被学习的权重，b_f表示偏置项，V表示显著区域特征，$V=\{v_1,v_2,\cdots,v_N\}$。

[0102] 如上述步骤S221-S224所述，可先获取每一个所述显著区域特征对应的第一线性表示，获取公式为$g_v=W_vV+b_v$；再获取每一个所述视觉常识特征对应的第二线性表示，获取公式为$g_c=W_cC+b_c$；g_v表示显著区域特征的线性表示，g_c表示视觉常识特征的线性表示，W_v、W_c表示需要被学习的权重，V表示显著区域特征，C表示视觉常识特征，b_c、b_v为偏置项，在通过第一线性表示与所述第二线性表示，计算拼接融合特征的线性表示施加影响，这样能够处理梯度爆炸和消失问题，使编码信息能够长时间深层不受阻碍地传播，在一定程度上能够防止常识信息的叠加和信息冗余。最后，通过$V_{acg}=\tanh(W_f f_{vc}+b_f)+V$所述拼接融合特征的模态间和模态内的常识性关系进行建模，得到ACG融合模型。

[0103] 在一个实施例中，将所述编码特征输入至Transformer解码器中完成训练，以搭建图像描述生成模型的步骤S3，包括：

[0104] S31、向Transformer解码器中的掩码自注意块输入标签信息，并将所述掩码自注意块作为第一子层，得到第一子层的第一特征信息；

[0105] S32、将所述第一特征信息与所述编码特征作为查询向量输入到Transformer解码

器中的交叉注意力块中，并将所述交叉注意力块作为第二子层，得到第二子层的第二特征信息；

[0106] S33、将所述第二特征信息输入到位置前馈网络进行非线性变换训练；

[0107] S34、返回到所述向Transformer解码器中的掩码自注意块输入标签信息的步骤，并对返回次数进行计数，得到返回总数；

[0108] S35、判断所述返回总数是否超过预设次数；

[0109] S36、若所述返回总数超过预设次数，判定所述第二特征信息训练完成，搭建图像描述生成模型。

[0110] 如上述步骤S31-S36所述，Transformer解码器由掩码自注意力块和交叉注意力块组成，掩码自注意力块的输入可以是一个标签信息，将其嵌入到特征矩阵 $Y=\{y_1,y_2,\ldots,y_n\},y_n\in\mathbf{R}^{n\times d_{\text{model}}}$ 中，n为输入标题长度，掩码自注意力块作为解码器的第一个子层，表示对标题词的自我注意，从而能够得到第一子层的第一特征信息，可通过公式：$S(Y)=f_{\text{mh-att}}(Y,Y,Y)$ 进行表示，其中：$Y\in\mathbf{R}^{n\times d_{\text{model}}}$ 是掩码标题特征；然后，将第一特征信息与编码特征 $\tilde{V}_{\text{acg}}$ 作为查询向量输入到交叉注意力块中(交叉注意力块作为解码器的第二个子层，表示对编码特征和标题特征的交叉注意)，得到第二特征信息，可通过公式：$S(Y,\tilde{V}_{\text{acg}})=f_{\text{mh-att}}(S(Y),\tilde{V}_{\text{acg}},\tilde{V}_{\text{acg}})$ 进行表示，最后，将 $S(Y,V_{acg})$ 输入到位置前馈网络，位置前馈网络由两个具有单一非线性的仿射变换组成，可以对特征进行非线性变换训练，其中，训练公式为：$F=FFN(S(Y,\tilde{V}_{\text{acg}}))$；FFN(•)表示位置前馈网络；线性词嵌入层以F为输入，将其转换为 d_v 维空间，其中 d_v 为词汇量。然后对每个单词执行softmax函数，预测字典中单词的概率。解码器包含N层，因此解码器重复相同的过程N次。在搭建完图像描述生成模型之后，可基于最小化交叉熵(XE) L_{XE} 对图像描述生成模型进行训练，涉及公式为：$L_{\text{XE}}(\theta)=-\sum_{t=1}^{T}\lg(p_\theta(y_t^*\mid y_{1:t-1}^*))$，其中，$y_{1:T}^*$ 表示设定的目标真值序列，θ为训练设置参数。

[0111] 本申请使用MSCOCO数据集来验证图像描述生成模型性能。MSCOCO数据集是当前图像描述任务的最大离线数据集，其包括82783个训练图像，40504个验证图像和40775个测试图像，每个图像标有5个标题。离线“Karpathy”数据分割用于离线性能比较，这种分割在图像描述工作中得到了广泛的应用，其中113,287张带有5个标题的图像进行训练，并用5000张图像用于验证，5000张图像用于测试。具体的实验环境在Windows操作系统下进行，基于Pytorch深度学习框架，该框架支持GPU运算；用于测试的环境为Python3.7；用于训练和测试的硬件配置为:Intel i7-9700 CPU 3.00GHz处理器，NVidia GeForce GTX 2080显卡。

[0112] 为了对所提出的模型方法进行定量的性能评价，实验采用了标准的客观量化评分方法，其中包括BLEU(BiLingual Evaluation Understudy，双语互评辅助工具)、ROUGE_L(Longest common subsequence based Recall-Oriented Understudy for Gisting Evaluation，基于最长公共子序列的面向记忆指标)、METEOR(Metric for Evaluation of Translation with Explicit Ordering，具有显式排序的翻译评价指标)、CIDEr(Consensus-based Image Description Evaluation，基于语义共识的图像描述评价指标)以及SPICE(Semantic Propositional Image Caption Evaluation，语义表达的图像描述

评价指标)等评价指标。其中BLEU为基于n-grams精确度的加权集合平均,该公式为

[0113] $$BLEU_N(c,s)=b(c,s)\exp\left(\sum_{n=1}^{N}\omega\lg cp_n(c,s)\right)$$

[0114] 例如:

[0115] Candidate:the the the the the the the.

[0116] Reference 1:The cat is on the mat.

[0117] Reference 2:There is a cat on the mat.

[0118] 其中,如果取n＝1,则一元组"the"在Candidate中出现的次数为7,在Reference 1中出现的次数为2,在Reference 2中出现的次数为1,则输出为2。N值取1、2、3、4,又可以分为BLEU-1、BLEU-2、BLEU-3、BLEU-4四个指标。

[0119] 参数设置方面,本申请在Transformer编码器和Transformer解码器中使用N＝3个相同的层。单词嵌入层的输出维数为512,输入的视觉特征也通过线性投影映射到512。前馈网络的内层维数为2048。多头注意采用h＝8个平行注意层。本申请使用Adam优化器训练,将beta1和beta2分别设置为0.9和0.999,epsilon设置为1e-6。在图像描述生成模型中设定单词时间步为20,为了生成更加合理的图像文字描述,本申请采用集束搜索Beam Search的方式,将beam size大小设置为3。为了增强图像描述生成模型的鲁棒性以及提高图像描述生成模型的训练速度,分别使用预先训练的Fatser R-CNN模型提取图像的显著区域特征,使用VC R-CNN模型提取图像对应视觉常识特征。训练中,本申请设置模型学习速率初始化为3e-4,输入批处理大小为10,每次训练5轮增加0.05的计划抽样概率,进行15轮交叉熵损失训练;随后使用SCST(Self-Critical Sequence Training for image caption generation)强化学习方法训练,学习率大小设置为1e-5,训练至30轮结束。

[0120] 本申请还提供一种基于Transformer结构的图像描述生成模型装置,包括:

[0121] 提取模块1,用于提取图像特征,其中,所述图像特征包括显著区域特征和视觉常识特征;

[0122] 分层输入模块2,用于将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中,以生成自适应融合的编码特征,其中,Transformer编码器包括多个分层,多个所述分层根据所述显著区域特征和视觉常识特征生成自适应常识门,所述自适应常识门用于对所述显著区域特征和所述视觉常识特征进行自适应融合;

[0123] 训练模块3,用于将所述编码特征输入至Transformer解码器中完成训练,以搭建图像描述生成模型;

[0124] 测试模块4,用于基于MSCOCO数据集对所述图像描述生成模型进行测试,以完成图像到语句的转化。

[0125] 在一个实施例中,所述提取模块1,包括:

[0126] 构建单元,用于基于Faster R-CNN构建图像的区域建议网络;

[0127] 引入单元,用于将所述区域建议网络引入深度神经网络模型中,得到基于深度神经网络的组合图像特征,将所述组合图像特征作为显著区域特征;

[0128] 提取单元,用于基于VCR-CNN提取图像边界框的坐标,其中,坐标包括多个;

[0129] 训练单元,用于将多个所述坐标输入卷积神经网络模型中进行训练,训练完成后得到视觉常识特征。

[0130] 在一个实施例中，所述引入单元，包括：

[0131] 第一图像获取子单元，用于基于所述区域建议网络获取多个不同批次的多个第一图像；

[0132] 第二图像获取子单元，用于对每一个批次的每一个所述第一图像进行短边缩放，得到每一个批次的短边缩放的多个第二图像；

[0133] 卷积池化单元，用于将每一个批次的多个所述第二图像传入卷积神经网络层中以对多个所述第二图像进行卷积和池化，以生成每一个批次的多个第二图像的组合图像特征。

[0134] 在一个实施例中，分层输入模块2，包括：

[0135] 拼接单元，用于对所述显著区域特征和所述视觉常识特征进行拼接，得到拼接融合特征；

[0136] 建模单元，用于根据所述拼接融合特征对图像模态间和模态内的常识性关系进行建模，得到ACG融合模型；

[0137] 训练单元，用于将所述拼接融合特征输入到所述ACG融合模型中进行训练，得到ACG输出特征；

[0138] 融合单元，用于将所述ACG输出特征分层输入到自注意力块中进行融合，得到多个层次的融合编码向量，其中，所述自注意力块包括多个，多个所述自注意力块进行模态内和跨模态的分层交互；

[0139] 处理单元，用于对所述融合编码向量进行残差和归一化处理，得到自适应融合的编码特征。

[0140] 在一个实施例中，所述拼接单元，包括：

[0141] 视觉特征向量子单元，用于基于所述视觉常识特征依次获取每一个所述视觉常识特征对应的视觉特征向量；

[0142] 显著区域向量子单元，用于基于所述显著区域特征依次获取每一个所述显著区域特征对应的显著区域向量；

[0143] 拼接子单元，用于根据所述视觉特征向量与所述显著区域向量对所述显著区域特征和所述视觉常识特征进行拼接，其中，拼接公式为：

[0144] $vc_i=[v_i,c_i]$；

[0145] $VC=\{vc_1,vc_2,\cdots,vc_N\}$；

[0146] 其中，所述$v_i\in R^d$，$c_i\in R^d$，$vc_i\in R^{2d}$，$v_i\in R^d$表示d维的第i个视觉特征向量，$c_i\in R^d$表示d维的第i个显著区域向量，$vc_i\in R^{2d}$表示2d维的显著区域特征和视觉常识特征的拼接向量；所述vc_i表示第i个显著区域特征和第i个视觉常识特征拼接；VC表示拼接融合特征。

[0147] 在一个实施例中，所述建模单元，包括：

[0148] 第一线性表示子单元，用于获取所述拼接融合特征中显著区域特征的第一线性表示；

[0149] 第二线性表示子单元，用于获取所述拼接融合特征中视觉常识特征的第二线性表示；

[0150] 计算线性表示子单元，用于根据所述第一线性表示与所述第二线性表示，计算拼接融合特征的线性表示施加影响，其中，计算公式为：

[0151] $f_{vc}=\mathrm{sigmoid}(g_v+g_c)*g_c$；

[0152] 其中，所述f_{vc}表示拼接融合特征的线性表示施加影响；g_v表示第一线性表示；g_c表示第二线性表示；

[0153] 建模子单元，用于根据所述线性表示施加影响对所述拼接融合特征的模态间和模态内的常识性关系进行建模，得到ACG融合模型，其中，建模过程为：

[0154] $V_{acg}=\tanh(W_f f_{vc}+b_f)+V$；

[0155] 其中，V_{acg}表示模态间和模态内的常识性关系，f_{vc}表示拼接融合特征的线性表示施加影响，W_f表示需要被学习的权重，b_f表示偏置项，V表示显著区域特征，$V=\{v_1,v_2,\cdots,v_N\}$。

[0156] 在一个实施例中，训练模块3，包括：

[0157] 输入标签信息单元，用于向Transformer解码器中的掩码自注意块输入标签信息，并将所述掩码自注意块作为第一子层，得到第一子层的第一特征信息；

[0158] 查询向量单元，用于将所述第一特征信息与所述编码特征作为查询向量输入到Transformer解码器中的交叉注意力块中，并将所述交叉注意力块作为第二子层，得到第二子层的第二特征信息；

[0159] 非线性变换训练单元，用于将所述第二特征信息输入到位置前馈网络进行非线性变换训练；

[0160] 返回单元，用于返回到所述向Transformer解码器中的掩码自注意块输入标签信息的步骤，并对返回次数进行计数，得到返回总数；

[0161] 判断单元，用于判断所述返回总数是否超过预设次数；

[0162] 判定单元，用于若所述返回总数超过预设次数，判定所述第二特征信息训练完成，搭建图像描述生成模型。

[0163] 如图3所示，本申请还提供了一种计算机设备，该计算机设备可以是服务器，其内部结构可以如图3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，该计算机设计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基于Transformer结构的图像描述生成模型方法的过程需要的所有数据。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基于Transformer结构的图像描述生成模型方法。

[0164] 本领域技术人员可以理解，图3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。

[0165] 本申请一实施例还提供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任意一个基于Transformer结构的图像描述生成模型方法。

[0166] 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与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，本申请所提供的和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。非易失性存储器可以包括只读存储器(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通过多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双速据率SDRAM(SSRSDRAM)、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。

[0167] 需要说明的是,在本文中,术语“包括”、“包含”或者其任何其它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、装置、物品或者方法不仅包括那些要素,而且还包括没有明确列出的其它要素,或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

[0168] 以上所述仅为本申请的优选实施例,并非因此限制本申请的专利范围,凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本申请的专利保护范围内。

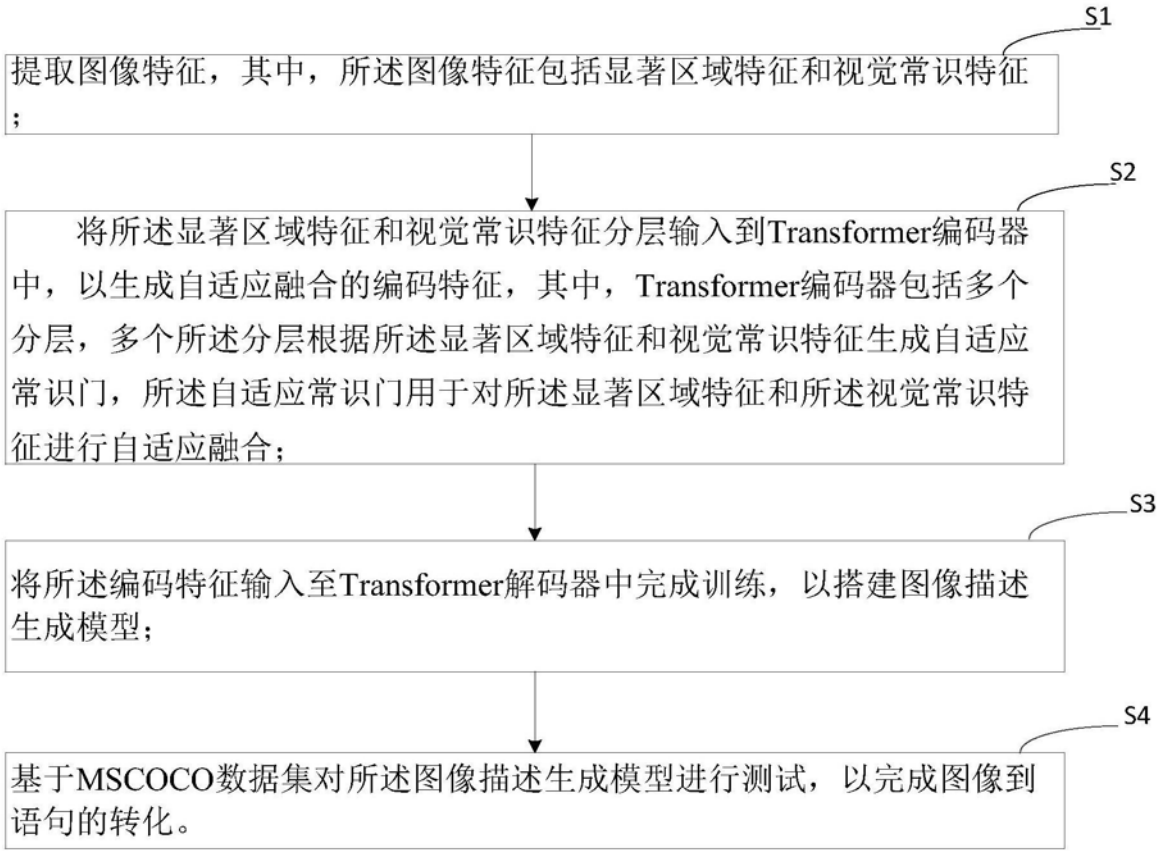

S1
提取图像特征，其中，所述图像特征包括显著区域特征和视觉常识特征；
S2
将所述显著区域特征和视觉常识特征分层输入到Transformer编码器中，以生成自适应融合的编码特征，其中，Transformer编码器包括多个分层，多个所述分层根据所述显著区域特征和视觉常识特征生成自适应常识门，所述自适应常识门用于对所述显著区域特征和所述视觉常识特征进行自适应融合；
S3
将所述编码特征输入至Transformer解码器中完成训练，以搭建图像描述生成模型；
S4
基于MSCOCO数据集对所述图像描述生成模型进行测试，以完成图像到语句的转化。

图1

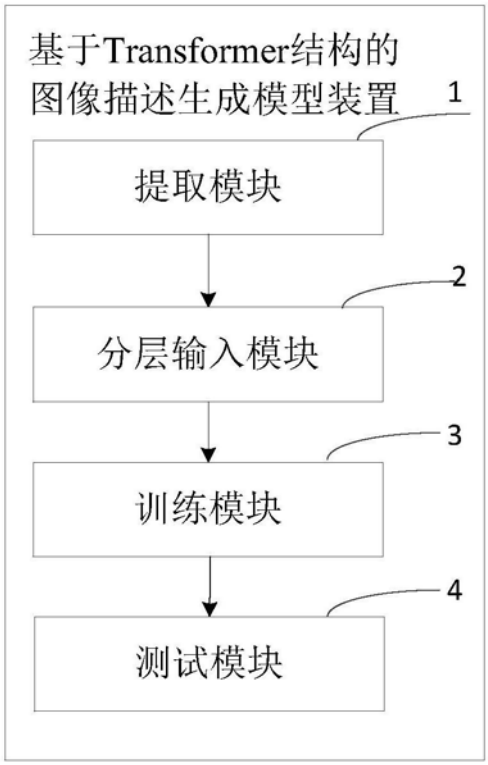


图2

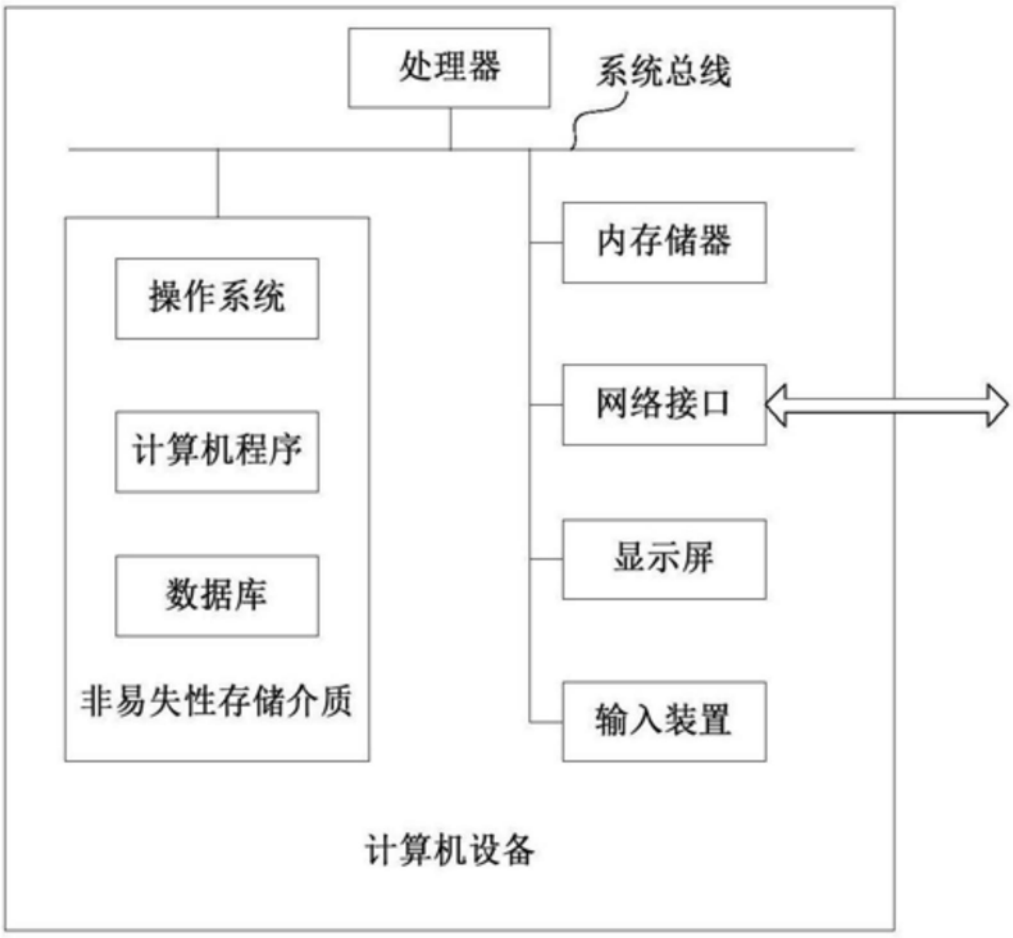


图3