(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114186568 B
(45)授权公告日 2022.08.02

(21)申请号 202111541714.X
(22)申请日 2021.12.16
(65)同一申请的已公布的文献号
申请公布号 CN 114186568 A
(43)申请公布日 2022.03.15
(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 刘云 石祎晖 冯方向 马占宇 王小捷
(74)专利代理机构 北京挺立专利事务所(普通合伙) 11265
专利代理师 叶盛 高福勇
(51)Int.Cl.
*G06F 40/30*(2020.01)
*G06N 3/04*(2006.01)
*G06N 3/08*(2006.01)
审查员 周永传

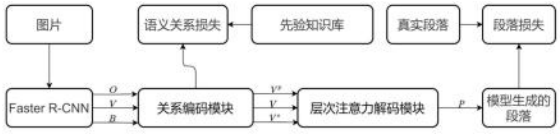
权利要求书3页 说明书10页 附图2页

(54)发明名称

一种基于关系编码和层次注意力机制的图像段落描述方法

(57)摘要

本发明公开了一种基于关系编码和层次注意力机制的图像段落描述方法,方法模型由关系编码模块和层次注意解码模块组成。关系编码模块通过两个编码器捕获编码空间关系信息和语义关系信息,其中语义关系编码时通过训练有监督的语义分类器来学习语义关系的先验知识。层次注意解码模块的层次注意力使用带有关系门和视觉门的层次注意力来动态的融合关系信息和物体区域特征,关系门用于在空间关系信息和语义关系信息之间切换,视觉门用于决定是否嵌入使用视觉信息,模型采用从粗粒度区域到细粒度的空间和语义关系的策略在段落生成过程中融合视觉信息。通过在斯坦福段落描述数据集上的大量实验表明,本发明方法在本领域的多个评价指标上显著优于现有方法。




CN 114186568 B

1.一种基于关系编码和层次注意力机制的图像段落描述方法,其特征在于,包括关系编码过程和层次注意力解码过程;

在关系编码过程中,输入区域特征V、区域位置B和区域类别O,通过空间关系编码器和语义关系编码器分别生成空间关系编码特征$V^P$和语义关系编码特征$V^S$,在语义关系编码时,从外部数据中收集语义物体关系对进行监督,通过训练有监督的语义关系分类器来学习语义关系编码的先验知识;

在层次注意解码过程中,使用两个LSTM和一个层次注意力动态融合关系信息和物体区域信息,层次注意力由具有关系门和视觉门的层次注意力组成,层次注意力分为一层区域注意力和一层关系注意力,区域注意力负责在生成当前单词时关注一个显著的物体,关系注意力由空间关系注意力和语义关系注意力组成,用于提取与被注意对象可能相关的关系信息;

空间关系编码过程的步骤为:

首先,根据物体框的几何结构得到相对坐标信息嵌入特征表示;给定两个物体框,$b_i=\{x_i,y_i,w_i,h_i\}$和$b_j=\{x_j,y_j,w_j\}$,它们的几何关系表示为四维向量λ(i,j),即:

$$\lambda(i,j)=\left(log\left(\frac{|x_i-x_j|}{w_i}\right),log\left(\frac{|y_i-y_j|}{h_i}\right),log\left(\frac{w_i}{w_j}\right),log\left(\frac{h_i}{h_j}\right)\right)$$

然后,使用一个线性层将λ(i,j)投影到一个高维空间中,该高维空间嵌入了两个物体框之间的相对坐标,如下式:

$E_b(i,j)=\mathrm{ReLU}(W_b\lambda(i,j)+b_b)$

其中$W_b\in R^{D_b\times 4}$和$b_b\in R^{D_b}$是可学习的参数;

通过相对坐标编码,空间关系信息编码$V^p=\{v_{ij}^p:v_{ij}^p\in R^{D_h}\}$由下式得到:

$v'_k=\mathrm{ReLU}(W_pv_k+b_p)$

$$v_{ij}^p=f_p(Concat(E_b(i,j),v_i^r,v_j^r))$$

其中,$W_p\in R^{D_h\times D_v}$和$b_p\in R^{D_h}$是可学习的权重和偏差,$v'_k$是物体区域特征向量$v_k$的低维投影,可学习的非线性函数$f_p(\cdot)$在实践中设置为一个两层的MLP,MLP的第一层和第二层设置相同,均具有一个ReLU激活函数、一个批量规范化和一个Dropout层;

语义关系编码过程的步骤为:

首先,两个物体$o_i$和$o_j$的$E_o(i,j)$的类别嵌入表示定义为:

$E_o(i,j)=\mathrm{ReLU}(W_o\mathrm{Concat}(W_go_i,W_go_j)+b_o)$

其中,$W_o\in R^{D_h\times 2D_g}$和$b_o\in R^{D_h}$是可学习的权重和偏差,$W_g\in R^{D_g\times D_o}$是一个固定的物体类别嵌入矩阵,该矩阵由GloVe向量初始化,在训练过程中保持不变;

然后,语义关系信息$V^s=\{v_{ij}^s:v_{ij}^s\in R^{D_h}\}$如下列公式所示:

$v''_k=\mathrm{ReLU}(W_sv_k+b_s)$

$$v_{ij}^s=f_s(Concat(E_b(i,j),E_o(i,j),v_i^{rr},v_j^{rr}))$$

其中,$W_s\in R^{D_h\times D_v}$和$b_s\in R^{D_h}$是可学习的权重和偏差;可学习的非线性函数$f_s(.)$在实践中设置为一个两层的MLP,MLP的第一层具有一个ReLU激活函数、一个批量规范化和一个Dropout层,第二层只具有单独的线性投影层;

语义关系分类器的步骤为：

首先，从Visual Genome数据集的视觉关系标注中收集语义关系三元组数据，两个物体$o_i$，$o_j$以及他们的语义关系$r_{ij} \in R^{D_r}$表示为语义关系三元组$(o_i, o_j, r_{ij})$；然后将编码为$v_{ij}^s$的语义关系输入一个线性层，以获得语义关系的类别分数，即：

$$\hat{r}_{ij} = W_r v_{ij}^s + b_r$$

其中$W_r \in R^{D_r \times D_h}$和$b_r \in R^{D_r}$是可学习的权重和偏差；

层次注意力的步骤为：

首先，通过如下公式获得物体区域注意力向量$a_o$：

$$a_{it} = w_{oa} tanh(W_{ov} v_i + W_{oh} h_t^1)$$

$\alpha_t = \mathrm{Softmax}(a_t)$

$$a_o = \sum_{i=1}^{N} \alpha_{it} v_i$$

其中，$W_{ov} \in R^{D_h \times D_v}$，$W_{oh} \in R^{D_h \times D_h}$和$w_{oa} \in R^{D_h}$是可学习的参数，$\alpha_{it}$表示每个对象特征$v_i$归一化注意权重；

然后，并行生成空间关系注意力向量$a_p$和语义关系注意力向量$a_s$；

空间关系注意力向量$a_p$生成方法为：在每一个时间步t中，通过采用空间注意力来生成空间关系注意力向量$a_p$：

$$p_{kt} = w_{pa} tanh(W_{pv} v_{gk}^p + W_{ph} h_t^1)$$

$\rho_t = \mathrm{Softmax}(p_t)$

$$a_p = \sum_{k=1}^{N} \rho_{kt} v_{gk}^p$$

其中，$W_{pv} \in R^{D_h \times D_h}$，$W_{ph} \in R^{D_h \times D_h}$和$w_{pa} \in R^{D_h}$是可学习的参数，$\rho_{kt}$表示空间关系特征$v_{gk}^p$的归一化注意权重；公式中$v_{gk}^p \in R^{D_h}$是对应物体区域g的第k个空间关系特征，通过获取第一层区域注意力对应物体的最大的注意权重$\alpha_{it}$来获得物体区域g；语义关系注意力向量$a_s$以与空间关系注意力向量$a_p$同样的方式得到；

$h_t^1$代表Attention LSTM的输出；

关系门$g_r$控制空间关系注意力向量$a_p$和语义关系注意力向量$a_s$，如下式所示：

$$g_r = \sigma(W_{rp} a_p + W_{rs} a_s + W_{rh} h_t^1)$$

其中，三个可学习的权重$W_{rp}$、$W_{rh}$和$W_{rs}$属于$R^{D_h \times D_h}$，$\sigma(\cdot)$表示sigmoid激活函数；

据此得到最终的关系注意力向量$a_r$，该向量表示同时包含了空间关系信息和语义关系信息，如下式所示：

$a_r = a_p \odot g_r + a_s \odot (1 - g_r)$

其中$\odot$表示逐个元素相乘的运算符号；

得到关系注意力向量$a_r$之后，将其输入一个线性层投影层，并将结果和物体区域注意力向量$a_o$相加并使用LayerNorm归一化，最终得到视觉上下文表示向量$a_v$，如下式所示：

$a_v = \mathrm{LayerNorm}(a_o + W_r a_r)$

其中，$W_r \in R^{D_h \times D_h}$是可学习的权重；

视觉门定义如下：

$$g_l = \sigma(W_{lx} x_t + W_{lh} h_t^1)$$

其中，$W_{lx} \in R^{D_h \times 3D_h}$和$W_{lh} \in R^{D_h \times D_h}$是可学习的权重，$x_t = Concat(h_{t-1}^2, \overline{v}, W_e y_{t-1})$是解码网络在每个时间步t时对Attention LSTM的输入；

据此得到了注意向量a，如下式所示：

$a = a_v \odot g_l + \tanh(m_t) \odot (1 - g_l)$

其中，$m_t$表示Attention LSTM的记忆单元在每个时间步t的输出；

最后通过将a与Attention LSTM的输出$h_t^1$拼接起来输入Language LSTM生成一个单词$y_t$，重复上述的过程直到生成结束符号为止，将生成的所有词拼接组成最终的段落即可。

2.根据权利要求1所述的基于关系编码和层次注意力机制的图像段落描述方法，其特征在于，对于重叠物体对，空间关系编码器通过拼接其视觉特征和相对位置坐标嵌入表示来获取空间关系编码的特征向量。

3.根据权利要求1所述的基于关系编码和层次注意力机制的图像段落描述方法，其特征在于，语义关系分类使用了多标签分类。

# 一种基于关系编码和层次注意力机制的图像段落描述方法

## 技术领域

[0001] 本发明涉及图像处理技术领域，尤其涉及一种基于关系编码和层次注意力机制的图像段落描述方法。

## 背景技术

[0002] 图像描述是为给定图像自动生成一个描述性句子的任务，也叫做图像单句描述。这项基本的跨模态任务可能有多种应用，如图像/视频检索、幼儿教育和帮助视力受损者理解图像内容。因此，这项任务引起了人工智能界的极大关注。

[0003] 在过去的几年中，许多研究在生成一个句子的图像描述任务上取得了令人印象深刻的进步。然而，由于一句话描述一幅图像的局限性，一句话对概括一幅图像中的各种细节通常是不够的，因为“一图胜千言”。为了解决一句话描述图像的局限性，Li Fei-Fei等人提出了图像段落描述的任务。一般来说，图像段落描述任务的目标是生成一个连贯的、细粒度的段落(通常包含四到六个句子)来描述给定的图像。

[0004] 以往关于图像段落描述的研究工作可分为两类:层次的方法和非层次的方法。层次的方法通过显式推断生成句子主题，然后通过句子主题生成句子组成段落。近年以来，人们提出了各种模型方法来改进图像段落描述任务，这些方法在很大程度上遵循编码器-解码器的框架。在最早期的工作中，Li Fei-Fei等人提出了一种层次的循环神经网络(Recurrent Neural Network,RNN)解码器来生成描述段落。该解码器由一个句子RNN和一个单词RNN组成，句子RNN负责生成句子的主题，单词RNN则根据已经生成的主题生成由单词组成的一句话，最后拼接所有的单词RNN生成的句子形成最终的描述段落。在之后的几年中，许多研究都提出了对层次解码结构的改进。另一方面，一些研究如把段落描述作为一个句子的词序列来进行生成段落，也实现了相似的性能和效果。

[0005] 然而，在以前的模型和方法中，图像中的单个物体通常由预训练的Faster R-CNN检测，之后表示为物体的区域特征。然后把图像中物体的区域特征输入后续的语言解码器来隐式地学习这些物体之间的关系，最终生成段落描述。因此，物体之间的关系对于生成准确、合理的描述非常有利，但这在之前的方法中没有得到充分的利用和编码。在图1中，给出了一个示例来显示用于图像段落描述的物体之间的细粒度关系(包括空间和语义关系)。在图中提到了多个物体，包括“beach”、“kite”、“water”、“man”和“clouds”。并给出了这些物体之间的空间关系(“kite-above-beach”和“kite-in-sky”)和语义关系(“man-flying-kite”和“man-standing on-beach”)。直观地说，物体之间的关系(包括空间关系和语义关系)可以丰富生成的段落描述的细节。

[0006] 在获得了物体间的关系信息之后，如何合理、有效的利用关系信息呢?一个简单的解决方案是将关系信息与物体特征结合(通过拼接或者是相加的方式)起来，然后将其放入语言解码器中，并以单层注意力的方式生成段落。然而，这种简单的融合方法存在着一个严重的问题。那就是关系信息和物体信息的融合纠缠可能会在生成段落时分散语言解码器的注意力，比如语言解码器需要去隐式地学习这些物体之间的关系。此外，这种简单的解决方

案与人类的层次认知过程不一致。具体来说，当一个人描述一幅图像时，他/她首先会注意到一个比较显著的物体，然后在描述这个物体时，他/她会进一步关注该物体与其他物体的关系，再进行描述，然后重复这个过程直到描述完成。在图1中，以第一句描述“A man is standing on the beach.”为例；我们首先注意到图像中有一个人，然后进一步注意到他“standing on”海滩上。这个例子表明，人类通过这种层次化的注意力机制，可以生成包含详细信息(比如关系信息)的句子并形成一个信息丰富的段落。因此，需要一种新的用于显式地利用更细粒度的空间和语义关系信息进行图像段落描述方法。

## 发明内容

[0007] 本发明针对上述技术问题，提供一种基于关系编码和层次注意力机制的图像段落描述方法。

[0008] 为了实现上述目的，本发明提供如下技术方案：

[0009] 一种基于关系编码和层次注意力机制的图像段落描述方法，包括关系编码过程和层次注意力解码过程；

[0010] 关系编码过程输入区域特征V、区域位置B和区域类别o，通过空间关系编码器和语义关系编码器分别生成空间关系编码特征$V^P$和语义关系编码特征$V^S$，在语义关系编码时，从外部数据中收集语义物体关系对进行监督，通过训练有监督的语义关系分类器来学习语义关系编码的先验知识；

[0011] 层次注意力解码过程使用两个LSTM和一个层次注意力动态融合关系信息和物体区域信息，层次注意力由具有关系门和视觉门的层次注意力组成，层次注意力分为一层区域注意力和一层关系注意力，区域注意力负责在生成当前单词时关注一个显著的物体，关系注意力由空间关系注意力和语义关系注意力组成，用于提取与被注意对象可能相关的关系信息。

[0012] 进一步地，对于重叠物体对，空间关系编码器通过拼接其视觉特征和相对位置坐标嵌入表示来获取空间关系编码的特征向量。

[0013] 进一步地，空间关系编码过程的步骤为：

[0014] 首先，根据物体框的几何结构得到相对坐标信息嵌入特征表示；给定两个物体框，$b_i=\{x_i,y_i,w_i,h_i\}$和$b_j=\{x_j,y_j,w_j,h_j\}$，它们的几何关系表示为四维向量$\lambda(i,j)$，即：

[0015] $$\lambda(i,j)=\left(log\left(\frac{|x_i-x_j|}{w_i}\right),log\left(\frac{|y_i-y_j|}{h_i}\right),log\left(\frac{w_i}{w_j}\right),log\left(\frac{h_i}{h_j}\right)\right)$$

[0016] 然后，使用一个线性层将$\lambda(i,j)$投影到一个高维空间中，该高维空间嵌入了两个物体框之间的相对坐标，如下式：

[0017] $E_b(i,j)=\mathrm{ReLU}(W_b\lambda(i,j)+b_b)$

[0018] 其中$W_b\in R^{D_b\times 4}$和$b_b\in R^{D_b}$是可学习的参数；

[0019] 通过相对坐标编码，空间关系信息编码$V^p=\{v_{ij}^p:v_{ij}^p\in R^{D_h}\}$由下式得到：

[0020] $v'_k=\mathrm{ReLU}(W_p v_k+b_p)$

[0021] $$v_{ij}^p=f_p(Concat(E_b(i,j),v_i',v_j'))$$

[0022] 其中，$W_p \in R^{D_h \times D_v}$和$b_p \in R^{D_h}$是可学习的权重，$v'_k$是物体区域特征向量$v_k$的低维投影，可学习的非线性函数$f_p(\cdot)$在实践中设置为一个两层的MLP，MLP的第一层和第二层设置相同，均具有一个ReLU激活函数、一个批量规范化和一个Dropout层。

[0023] 进一步地，语义关系编码过程的步骤为：

[0024] 首先，两个物体$o_i$和$o_j$的$E_o(i,j)$的类别嵌入表示定义为：

[0025] $E_o(i,j)=\mathrm{ReLU}(W_o\mathrm{Concat}(W_g o_i, W_g o_j)+b_o)$

[0026] 其中，$W_o \in R^{D_h \times 2D_g}$和$b_o \in R^{D_h}$是可学习的权重和偏差，$W_g \in R^{D_g \times D_o}$是一个固定的物体类别嵌入矩阵，该矩阵由GloVE向量初始化，在训练过程中保持不变；

[0027] 然后，语义关系信息$V^s = \{v_{ij}^s : v_{ij}^s \in R^{D_h}\}$如下列公式所示：

[0028] $v''_k=\mathrm{ReLU}(W_s v_k+b_s)$

[0029] $$v_{ij}^s = f_s(Concat(E_b(i,j), E_o(i,j), v_i'', v_j''))$$

[0030] 其中，$W_s \in R^{D_h \times D_v}$和$b_s \in R^{D_h}$是可学习的权重和偏差；可学习的非线性函数$f_p(\cdot)$在实践中设置为一个两层的MLP，MLP的第一层具有一个ReLU激活函数、一个批量规范化和一个Dropout层，第二层只具有单独的线性投影层。

[0031] 进一步地，语义关系分类器的步骤为：首先，从Visual Genome数据集的视觉关系标注中收集语义关系三元组数据，两个物体$o_i$，$o_j$以及他们的语义关系$r_{ij} \in R^{D_r}$表示为语义关系三元组$(o_i, o_j, r_{ij})$；然后将编码为$v_{ij}^s$的语义关系输入一个线性层，以获得语义关系的类别分数，即：

[0032] $$\hat{r}_{ij} = W_r v_{ij}^s + b_r$$

[0033] 其中$W_r \in R^{D_r \times D_h}$和$b_o \in R^{D_r}$是可学习的权重和偏差。

[0034] 进一步地，语义关系分类使用了多标签分类。

[0035] 进一步地，层次注意力的步骤为：

[0036] 首先，通过如下公式获得物体区域注意力$a_o$：

[0037] $$a_{it} = w_{oa} tanh(W_{ov} v_i + W_{oh} h_t^1)$$

[0038] $\alpha_t=\mathrm{Softmax}(a_t)$

[0039] $$a_o = \sum_{i=1}^{N} \alpha_{it} v_i$$

[0040] 其中，$W_{ov} \in R^{D_h \times D_v}$，$W_{oh} \in R^{D_h \times D_h}$和$w_{oa} \in R^{D_h}$是可学习的参数，$\alpha_{it}$表示每个对象特征$v_i$归一化注意力注意权重；

[0041] 然后，并行生成空间关系上下文向量$a_p$和语义关系上下文向量$a_s$。

[0042] 进一步地，空间关系上下文向量$a_p$生成方法为：在每一个时间步t中，通过采用空间注意力来生成空间关系注意力向量$a_p$：

[0043] $$a_{it} = w_{pa} tanh(W_{pv} v_{gi}^p + W_{ph} h_t^1)$$

[0044] $\alpha_t = \mathrm{Softmax}(a_t)$

[0045] $$a_p = \sum_{i=1}^{N} \alpha_{it} v_{gi}^p$$

[0046] 其中，$W_{pv} \in R^{D_h \times D_h}$，$W_{ph} \in R^{D_h \times D_h}$和$w_{pa} \in R^{D_h}$是可学习的参数，$a_{it}$表示空间关系特征$v_{gi}^p$的归一化注意权重；公式中$v_{gi}^p \in R^{D_h}$是对应物体区域g的第i个空间关系特征，通过获取第一层区域注意力对应物体的最大注意权重$\alpha_{it}$来获得物体区域g；语义关系注意力向量$a_s$以与空间关系注意力向量$a_p$同样的方式得到。

[0047] 进一步地，关系门$g_r$控制空间关系注意力向量$a_p$和语义关系注意力向量$a_s$，如下式所示：

[0048] $$g_r = \sigma\left(W_{rp} a_p + W_{rs} a_s + W_{rh} h_t^1\right)$$

[0049] 其中，三个可学习的权重$W_{rp}$、$W_{rh}$和$W_{rs}$属于$R^{D_h \times D_h}$，σ(•)表示sigmoid激活函数；

[0050] 据此得到最终的关系注意力向量$a_r$，该向量表示同时包含了空间关系信息和语义关系信息，如下式所示：

[0051] $a_r = a_p \odot g_r + a_s \odot (1 - g_r)$

[0052] 其中⊙表示逐个元素相乘的运算符号；

[0053] 得到关系注意力向量$a_r$之后，将其输入一个线性层投影层，并将结果和到物体区域注意力向量$a_o$相加并使用LayerNorm归一化，最终得到视觉上下文表示向量$a_v$，如下式所示：

[0054] $a_v = \mathrm{LayerNorm}(a_o + W_r(a_r))$

[0055] 其中，$W_r \in R^{D_h \times D_h}$是可学习的权重。

[0056] 进一步地，视觉门定义如下：

[0057] $$g_l = \sigma(W_{lx} x_t + W_{lh} h_t^1)$$

[0058] 其中，$W_{lx} \in R^{D_h \times 3D_h}$和$W_{lh} \in R^{D_h \times D_h}$是可学习的权重，$x_t = Concat(h_{t-1}^2, \bar{v}, W_e y_{t-1})$是解码网络在每个时间步t时对Attention LSTM的输入；

[0059] 据此得到了注意向量a，如下式所示：

[0060] $a = a_v \odot g_l + \tanh(m_t) \odot (1 - g_l)$

[0061] 其中，$m_t$表示Attention LSTM的记忆单元在每个时间步t的输出；

[0062] 最后通过将a与Attention LSTM的输出$h_t^1$拼接起来输入Language LSTM生成一个单词$y_t$，重复上述的过程直到生成结束符号为止，将生成的所有词拼接组成最终的段落即可。

[0063] 与现有技术相比，本发明的有益效果为：

[0064] 本发明提供的基于关系编码和层次注意力机制的图像段落描述方法(DualRel)，是一种用于图像段落字幕的任务的新方法，DualRel模型的动机是有效地利用图像中存在的细粒度的空间和语义关系。为此，DualRel模型由关系编码模块和层次注意解码模块组成。关系编码模块通过两个编码器捕获图像中物体之间的空间关系信息和语义关系信息，

利用细粒度的空间和语义关系信息,在编码过程中,语义关系编码时我们通过训练有监督的语义分类器来学习和语义关系有关的先验知识。层次注意解码模块以Top-Down注意力网络为原型。层次注意力使用带有关系门和视觉门的层次注意力来动态的融合关系信息和物体区域特征,我们设计的关系门用于在两种关系信息(空间关系信息和语义关系信息)之间切换,设计的视觉门用于决定是否嵌入使用视觉信息,采用从粗粒度区域到细粒度的空间和语义关系的策略在段落生成过程中融合视觉信息。通过在斯坦福段落描述数据集(Stanford Benchmark Dataset)上的大量实验表明,本发明的方法在本领域的多个评价指标上显著优于现有的方法。

## 附图说明

[0065] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例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中记载的一些实施例,对于本领域普通技术人员来讲,还可以根据这些附图获得其他的附图。

[0066] 图1为图像描述段落中的空间和语义关系展示,物体之间的空间关系如“kite-above-beach”和语义关系如“man-standing on-beach;

[0067] 图2为本发明实施例提供的DualRel模型的架构图;

[0068] 图3为本发明实施例提供的关系编码模块的架构图,包含空间编码器,语义编码器和语义关系分类器。

[0069] 图4为本发明实施例提供的层次注意力解码模块的架构图,包含区域注意力,两个关系注意力和两个门控(关系门和视觉门)。

## 具体实施方式

[0070] 为了使本领域的技术人员更好地理解本发明的技术方案,下面将结合附图和实施例对本发明作进一步的详细介绍。

[0071] 本发明的基于关系编码和层次注意力机制的图像段落描述方法(DualRel),DualRel模型详情如图2所示。我们的DualRel模型包含两个主要模块,一个关系编码模块和一个层次注意力解码模块。关系编码模块输入区域特征V,区域位置B和区域类别O,通过空间关系编码器和语义关系编码器分别生成空间关系编码特征$V^P$和语义关系编码特征$V^S$,此外为了监督模型学习到有关语义关系的先验知识,我们提出了一个新颖的语义关系分类损失,该损失用于前期帮助模型学习到通用的先验语义关系信息。为了更好的利用学习到的特征V,$V^P$和$V^S$,让他们在解码过程中更好的交互融合,我们提出了一个层次注意力解码模块,该模块通过使用层次注意力和门控机制来生成最终的段落P。接下来,我们将详细的介绍关系编码模块和层次注意解码模块。

[0072] 对于图像段落描述来说,我们的目标是为任何给定的图像I生成一个段落$P=\{y_1,\cdots,y_T\}$,其中T表示生成描述的长度。本文中图像特征使用预训练的Faster R-CNN提取。使用$O=\{o_1,\cdots,o_N\}$表示检测到的N个物体,检测到的物体个数取决于输入图像。让$V=\{v_1,\cdots,v_N\},v_k\in R^{D_v}$作为它们的视觉特征表示,而$B=\{b_1,\cdots,b_N\}$,$b_i=\{x_i,y_i,w_i,h_i\}\in R^4$作为它们物体边界框。其中(x,y)表示物体框的中心坐标,(w,h)表示物体框的宽度和高度。

此外,图像的全局表示$\overline{v}=\frac{1}{N}\sum_{i=1}^{N}v_i$包含了总体的图像特征。

[0073] 关系编码模块概述如图3所示。

[0074] 空间关系编码器(Spatial Relation Encoder):如前所述所述,为了生成详细的段落描述,我们需要获取物体之间的空间关系信息(例如“above”和“on”)。我们观察到,许多描述场景中物体的句子通常只包含附近物体的空间位置关系。因此,在本文中我们只考虑一个物体与另一个物体重叠的情况来进行空间关系信息的编码。对于重叠物体对,我们通过拼接其视觉特征和相对位置坐标嵌入表示来获取空间关系编码的特征向量。

[0075] 具体地,空间关系编码器的步骤为:

[0076] 首先,根据物体框的几何结构得到相对坐标信息嵌入特征表示;给定两个物体框,$b_i=\{x_i,y_i,w_i,h_i\}$和$b_j=\{x_j,y_j,w_j,h_j\}$,它们的几何关系表示为四维向量$\lambda(i,j)$,即:

[0077] $$\lambda(i,j)=\left(log\left(\frac{|x_i-x_j|}{w_i}\right),log\left(\frac{|y_i-y_j|}{h_i}\right),log\left(\frac{w_i}{w_j}\right),log\left(\frac{h_i}{h_j}\right)\right)$$

[0078] 然后,使用一个线性层将$\lambda(i,j)$投影到一个高维空间中,该高维空间嵌入了两个物体框之间的相对坐标,如下式:

[0079] $E_b(i,j)=\mathrm{ReLU}(W_b\lambda(i,j)+b_b)$

[0080] 其中$W_b\in R^{D_b\times 4}$和$b_b\in R^{D_b}$是可学习的参数;

[0081] 通过相对坐标编码,空间关系信息编码$V^p=\{v_{ij}^p:v_{ij}^p\in R^{D_h}\}$由下式得到:

[0082] $v'_k=\mathrm{ReLU}(W_pv_k+b_p)$

[0083] $$v_{ij}^p=f_p(Concat(E_b(i,j),v_i',v_j'))$$

[0084] 其中,$W_p\in R^{D_h\times D_v}$和$b_p\in R^{D_h}$是可学习的权重,$v'_k$是物体区域特征向量$v_k$的低维投影,可学习的非线性函数$f_p(\cdot)$在实践中设置为一个两层的MLP(Multi-layer Perceptron),MLP的第一层和第二层设置相同,均具有一个ReLU激活函数、一个批量规范化和一个Dropout层。

[0085] 语义关系编码器(Semantic Relation Encoder):语义关系编码器用于编码两个物体之间另一种类型的关系信息(例如,“flying”和“eating”),这对于生成描述至关重要。如前文所述,与空间关系不同,语义关系需要一定的先验知识学习才能推断出来。并且我们观察到,在对象类别和它们的语义关系之间存在着很强的相关性,比如“human”和“bike”之间的关系大概率是“riding”或者“push”,而不会是“eating”或者“flying”这些关系,因此在编码物体之间的语义关系时,我们会显式的加入两个物体的类别信息O。

[0086] 具体地,语义关系编码过程的步骤为:

[0087] 首先,两个物体$o_i$和$o_j$的$E_o(i,j)$的类别嵌入表示定义为:

[0088] $E_o(i,j)=\mathrm{ReLU}(W_o\mathrm{Concat}(W_go_i,W_go_j)+b_o)$

[0089] 其中,$W_o\in R^{D_h\times 2D_g}$和$b_o\in R^{D_h}$是可学习的权重和偏差,$W_g\in R^{D_g\times D_o}$是一个固定的物体类别嵌入矩阵,该矩阵由GloVE向量初始化,在训练过程中保持不变;

[0090] 然后,语义关系信息$V^s=\{v_{ij}^s:v_{ij}^s\in R^{D_h}\}$如下列公式所示:

[0091] $v''_k = \mathrm{ReLU}(W_s v_k + b_s)$

[0092] $v_{ij}^s = f_s(Concat(E_b(i,j), E_o(i,j), v_i'', v_j''))$

[0093] 其中，$W_s \in R^{D_h \times D_v}$和$b_s \in R^{D_h}$是可学习的权重和偏差;可学习的非线性函数$f_p(\cdot)$在实践中设置为一个两层的MLP,MLP的第一层具有一个ReLU激活函数、一个批量规范化和一个Dropout层,第二层只具有单独的线性投影层。

[0094] 语义关系分类器(Semantic Relation Classifier):对于语义关系编码器来说,直接从段落标注中直接学习语义关系是困难的,因为语义关系学习需要大量的先验知识监督,而段落的解码生成过程距离语义关系编码器太远,可能无法在模型早期训练时实现有效的学习。

[0095] 因此我们设计了一个语义关系分类器,利用先验知识对语义关系编码器进行显式监督。

[0096] 具体地,语义关系分类器的步骤为:首先,从Visual Genome数据集的视觉关系标注中收集语义关系三元组数据,两个物体$o_i$,$o_j$以及他们的语义关系$r_{ij} \in R^{D_r}$表示为语义关系三元组($o_i$,$o_j$,$r_{ij}$);然后将编码为$v_{ij}^s$的语义关系输入一个线性层,以获得语义关系的类别分数,即:

[0097] $\hat{r}_{ij} = W_r v_{ij}^s + b_r$

[0098] 其中$W_r \in R^{D_r \times D_h}$和$b_o \in R^{D_r}$是可学习的权重和偏差。

[0099] 值得注意的是语义关系分类使用了多标签分类任务,因为两个物体之间可能存在多个关系,因为我们没有真实的两个物体之间关系的标注。

[0100] 层次注意力解码模块(Hierarchical Attention Decoding Module):如前文所述,我们在关系编码模块中提取了空间关系特征VP和语义关系特征VS,并提取对象区域特征V。为了融合这三个特征,生成包含更多关系的段落。基于人类层次的认知过程,我们提出了层次注意解码模块。具体来说,当人类描述一个图像时,我们首先会观察并注意到一个显著的物体,然后在描述这个物体的过程中,我们会进一步关注这个物体与其他物体之间的关系信息(包括空间和语义关系信息),从而生成一个信息性和描述性的段落。层次注意力解码模块如图4所示。我们的解码模块基于Top-Down注意力网络设计。我们设计了具有关系门和视觉门的空间和语义关系注意力的层次注意力模块来替换原模型的注意力模块。接下来我们将详细描述我们设计的层次注意力和门控机制的详细情况。

[0101] 层次注意力(Hierarchical Attention):Top-Down注意力网络包括一个Attention LSTM、一个Language LSTM和一个注意力模块。在生成段落期间的每个时间步t时,可将其形式化为:

[0102] $h_t^1 = LSTM_{att}(Concat(h_{t-1}^2, \bar{v}, W_e y_{t-1}), h_{t-1}^1)$

[0103] $a = Attention(V, h_t^1)$

[0104] $h_t^2 = LSTM_{lang}(Concat(a, h_t^1), h_{t-1}^2)$

[0105] 其中，$h_t^1 \in R^{D_h}$是Attention LSTM的输出，$W_e \in R^{D_e \times D_w}$是词汇表的词嵌入矩阵，$y_{t-1}$

是输入单词在每个时间步时间t的一个独热编码。$a \in R^{D_h}$是注意力向量，$h_t^2 \in R^{D_h}$是Language LSTM的输出。

[0106] 具体地，层次注意力的步骤为：

[0107] 首先，通过如下公式获得物体区域注意力$a_o$：

[0108] $$a_{it} = w_{oa} tanh(W_{ov} v_i + W_{oh} h_t^1)$$

[0109] $\alpha_t = Softmax(a_t)$

[0110] $$a_o = \sum_{i=1}^{N} \alpha_{it} v_i$$

[0111] 其中，$W_{ov} \in R^{D_h \times D_v}$，$W_{oh} \in R^{D_h \times D_h}$和$w_{oa} \in R^{D_h}$是可学习的参数，$\alpha_{it}$表示每个对象特征$v_i$归一化注意力注意权重；

[0112] 然后，并行生成空间关系上下文向量$a_p$和语义关系上下文向量$a_s$。

[0113] 空间关系上下文向量$a_p$生成方法为：在每一个时间步t中，通过采用空间注意力来生成空间关系注意力向量$a_p$：

[0114] $$a_{it} = w_{pa} tanh(W_{pv} v_{gi}^p + W_{ph} h_t^1)$$

[0115] $\alpha_t = Softmax(a_t)$

[0116] $$a_p = \sum_{i=1}^{N} \alpha_{it} v_{gi}^p$$

[0117] 其中，$W_{pv} \in R^{D_h \times D_h}$，$W_{ph} \in R^{D_h \times D_h}$和$w_{pa} \in R^{D_h}$是可学习的参数，$\alpha_{it}$表示空间关系特征$v_{gi}^p$的归一化注意权重；公式中$v_{gi}^p \in R^{D_h}$是对应物体区域g的第i个空间关系特征，通过获取第一层区域注意力对应物体的最大注意权重$\alpha_{it}$来获得物体区域g；以同样的方式，我们可以得到语义关系注意力向量$a_s$。

[0118] 关系门(Relational Gate)：在前文中我们通过层次的注意力机制获取了空间关系注意力向量$a_p$和语义关系注意力向量$a_s$。为了控制这两类关系信息在解码过程中如何融合使用，我们设计了一个关系门$g_r$来控制两种信息如何使用，具体地：

[0119] 关系门$g_r$控制空间关系注意力向量$a_p$和语义关系注意力向量$a_s$，如下式所示：

[0120] $$g_r = \sigma(W_{rp} a_p + W_{rs} a_s + W_{rh} h_t^1)$$

[0121] 其中，三个可学习的权重$W_{rp}$、$W_{rh}$和$W_{rs}$属于$R^{D_h \times D_h}$，$\sigma(\cdot)$表示sigmoid激活函数；

[0122] 据此得到最终的关系注意力向量$a_r$，该向量表示同时包含了空间关系信息和语义关系信息，如下式所示：

[0123] $a_r = a_p \odot g_r + a_s \odot (1 - g_r)$

[0124] 其中$\odot$表示逐个元素相乘的运算符号；

[0125] 得到关系注意力向量$a_r$之后，将其输入一个线性层投影层，并将结果和到物体区域注意力向量$a_o$相加并使用LayerNorm归一化，最终得到视觉上下文表示向量$a_v$，如下式所示：

[0126] $a_v = \mathrm{LayerNorm}(a_o + W_r(a_r))$

[0127] 其中，$W_r \in R^{D_h \times D_h}$是可学习的权重。

[0128] 视觉门(Visual Gate)：我们定义了一个视觉门来决定在解码时使用视觉信息还是使用语言上下文信息。直观来说，解码器在生成一些词的时比如“the”和“is”，可能只需要需要很少的视觉信息来生成这些单词。视觉门定义如下：

[0129] $$g_l = \sigma(W_{lx} x_t + W_{lh} h_t^1)$$

[0130] 其中，$W_{lx} \in R^{D_h \times 3D_h}$和$W_{lh} \in R^{D_h \times D_h}$是可学习的权重，$x_t = Concat(h_{t-1}^2, \bar{v}, W_e y_{t-1})$是解码网络在每个时间步t时对Attention LSTM的输入；

[0131] 据此得到了注意向量a，如下式所示：

[0132] $a = a_v \odot g_l + \tanh(m_t) \odot (1 - g_l)$

[0133] 其中，$m_t$表示Attention LSTM的记忆单元在每个时间步t的输出；

[0134] 最后通过将a与Attention LSTM的输出$h_t^1$拼接起来输入Language LSTM生成一个单词$y_t$，重复上述的过程直到生成结束符号为止，将生成的所有词拼接组成最终的段落即可。

[0135] 此外，关于随时函数(Loss Function)的说明如下：

[0136] 语义关系分类损失(Semantic Relation Classification Loss)：语义关系分类损失的目的是鼓励模型利用先验知识学习语义关系编码。我们应用了多标签分类损失函数，即：

[0137] $$L_R = log\left(1 + \sum_{p \in \Omega_{neg}} e^{\widehat{r_{ij}^p}}\right) + log\left(1 + \sum_{q \in \Omega_{pos}} e^{-\widehat{r_{ij}^q}}\right)$$

[0138] 其中$\widehat{r_{ij}^p}$是语义关系分类器输出的某一个语义关系的类别分数。集合$\Omega_{neg}$表示两个物体$o_i$和$o_j$没有特定的某一类语义关系t(即$r_{ij}^t$)，集合$\Omega_{pos}$表示两个对象具有某一类特定的语义关系。

[0139] 词级损失(Word-level Loss)：给定一幅图像和真实标注的段落对(I,P)，我们通过最大化和真实标注段落P的相似性来训练DualRel模型，这等价于最小化交叉熵(XE)损失：

[0140] $$L_{XE}(\theta) = -\sum_{t=1}^{T} log(p_\theta(y_t | y_{1:t-1}))$$

[0141] 总体损失(Total Loss)：最终损失函数定义为语义关系分类损失和词级损失的线性组合。具体而言，总体损失L定义如下：

[0142] $L = \zeta L_R + \eta L_{XE}$

[0143] 其中ζ和η是不同损失的权重。该权重通过实验确定，在模型实现细节里我们会进一步介绍这两个权重的取值。

[0144] SCST(Self-critical Sequence Training)：为了提升模型的效果，我们进一步使用自我批评序列训练(SCST)的方式优化了我们的模型。指标的期望梯度计算如下：

[0145] $$\nabla_\theta L(\theta) = -E_{w^s \sim p_\theta}\left[\left(r(w^s) - r(w^g)\right)\nabla_\theta \log p_\theta(w^s)\right]$$

[0146] 其中，$w^s$和$w^g$分别表示依据概率采样的段落和贪婪地采样段落。r(•)表示来自段落评价指标的奖励，$p_\theta$表示DualRel模型的参数。此外，我们采用了两种类型的奖励的SCST训练模型。其中一种是只使用CIDEr，这用于公平比较。另一种是CIDEr,METEOR和BLEU-4三种指标的混合训练模型。

[0147] 综上，我们提出了一个新颖的名为DualRel的新模型，用于显式地利用更细粒度的空间和语义关系信息进行图像段落描述。

[0148] 首先，我们设计了一个关系编码模块，由空间关系编码器和语义关系编码器两个部分组成。空间关系编码器强调重叠物体之间的空间位置关系的编码。而语义关系编码器则用于编码物体之间的语义关系信息。为了有效地学习语义关系相关的先验知识，我们提出使用从外部数据中收集语义物体关系对，然后构造语义关系分类器来显式的监督模型学习语义关系的先验知识。

[0149] 其次，我们设计了一个层次注意解码模块，该模块使用两个LSTM和一个层次注意力动态融合关系信息和物体区域信息。层次注意力分为一层区域注意力和一层关系注意力，区域注意力负责在生成当前单词时关注一个显著的物体。关系注意力由空间关系注意力和语义关系注意力组成，用于提取与被注意对象可能相关的关系信息。

[0150] 此外，关系门控制所需的关系信息类型(语义关系还是空间关系信息)。视觉门决定输出特征是依赖于视觉信息还是语言上下文信息。

[0151] 我们的主要贡献如下：

[0152] 1、我们提出了DualRel模型用于图像段落描述，该模型由关系编码模块和层次注意解码模块组成。关系编码模块通过两个编码器编码空间和语义关系信息。在编码过程中，语义关系编码时我们通过训练有监督的语义分类器来学习和语义关系有关的先验知识。

[0153] 2、我们设计了一个层次化的注意力解码模块来动态地融合利用细粒度的关系信息和物体区域信息。层次注意力由具有关系门和视觉门的层次注意力组成。

[0154] 3、我们在斯坦福段落描述数据集(Stanford Benchmark Dataset)上进行了广泛的实验。我们采用了七种流行的评估指标，包括BLEU-1,BLEU-2,BLEU-3,BLEU-4,METEOR和CIDEr以及BERTScore的F值指标。我们的模型在BLEU-1,BLEU-2,BLEU-3,BLEU-4上分别实现了45.30,28.91,18.46,11.30的分数，在CIDEr值上实现了34.02的分数，达到了84.37的$F_{BERT}$(idf)BERTScore分数，现有基础方法的BLEU-1,BLEU-2,BLEU-3,BLEU-4,CIDEr和$F_{BERT}$(idf)分数分别为43.54,27.44,17.33,10.58,30.64和83.85，这些实验结果表明，我们提出的DualRel在本领域的多个评价指标上显著优于现有的方法，并且具有实用性和创新性。

[0155] 以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换，但这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

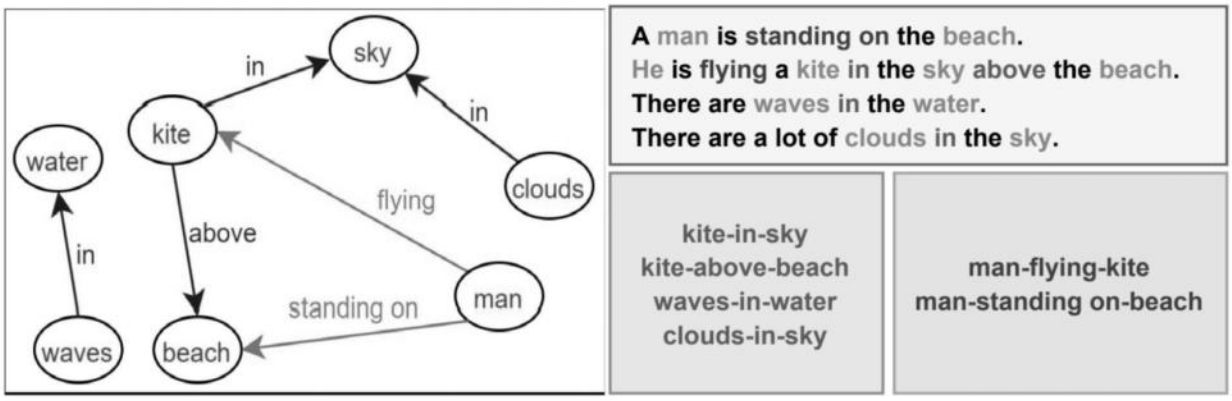

sky
in
kite
in
water
clouds
flying
above
in
man
standing on
waves
beach
A man is standing on the beach.
He is flying a kite in the sky above the beach.
There are waves in the water.
There are a lot of clouds in the sky.
kite-in-sky
kite-above-beach
waves-in-water
clouds-in-sky
man-flying-kite
man-standing on-beach

图1

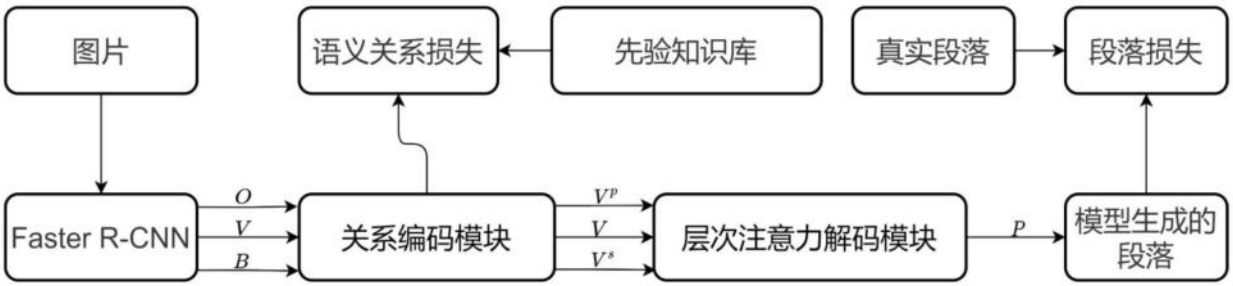

图片
语义关系损失
先验知识库
真实段落
段落损失
Faster R-CNN
O
V
B
关系编码模块
$V^p$
V
$V^s$
层次注意力解码模块
P
模型生成的段落

图2

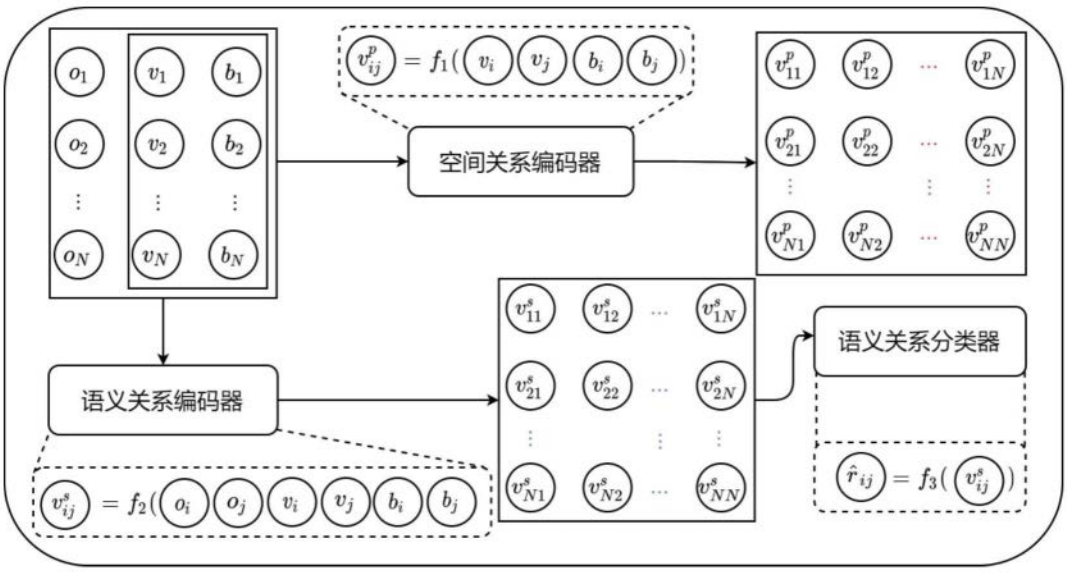

$o_1$ $o_2$ $o_N$
$v_1$ $v_2$ $v_N$
$b_1$ $b_2$ $b_N$
$v_{ij}^p = f_1(v_i\ v_j\ b_i\ b_j)$
空间关系编码器
$v_{11}^p$ $v_{12}^p$ $v_{1N}^p$
$v_{21}^p$ $v_{22}^p$ $v_{2N}^p$
$v_{N1}^p$ $v_{N2}^p$ $v_{NN}^p$
语义关系编码器
$v_{11}^s$ $v_{12}^s$ $v_{1N}^s$
$v_{21}^s$ $v_{22}^s$ $v_{2N}^s$
$v_{N1}^s$ $v_{N2}^s$ $v_{NN}^s$
语义关系分类器
$v_{ij}^s = f_2(o_i\ o_j\ v_i\ v_j\ b_i\ b_j)$
$\hat{r}_{ij} = f_3(v_{ij}^s)$

图3

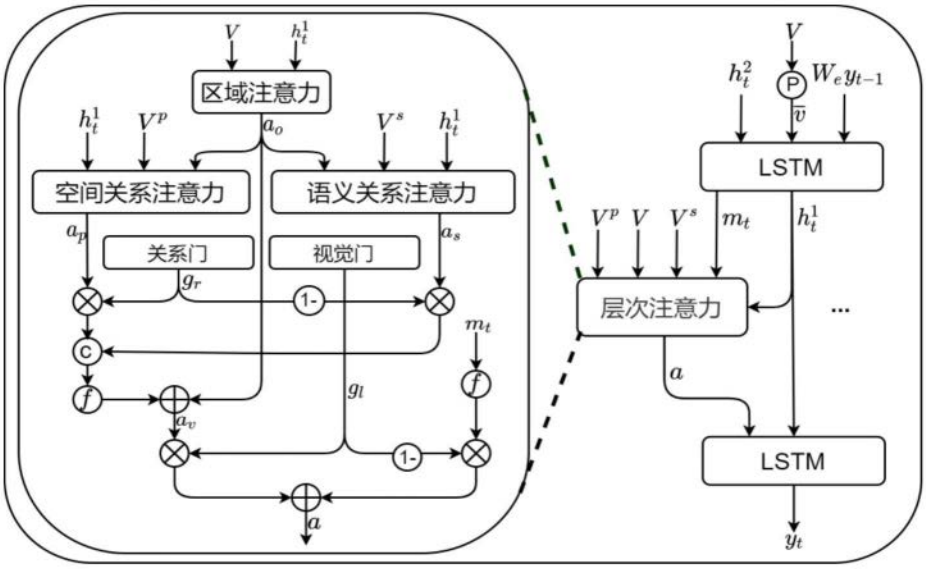



图4