(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117609536 A
(43)申请公布日 2024.02.27

(21)申请号 202311632626.X

(22)申请日 2023.12.01

(71)申请人 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 陆明聪 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 高福勇

(51)Int.Cl.
G06F 16/58(2019.01)
G06F 16/583(2019.01)
G06V 10/44(2022.01)
G06V 10/764(2022.01)
G06V 10/82(2022.01)
G06F 16/55(2019.01)
G06F 16/33(2019.01)
G06N 3/0455(2023.01)
G06N 3/048(2023.01)
G06N 3/08(2023.01)
G06N 5/04(2023.01)

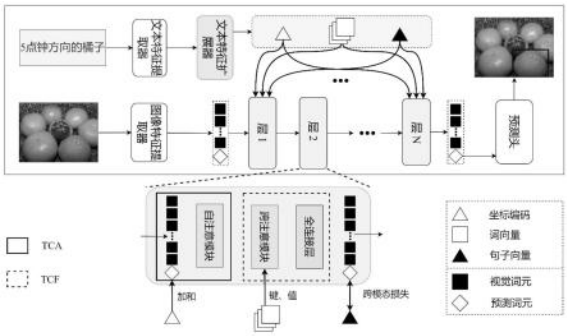
权利要求书4页 说明书9页 附图2页

(54)发明名称

基于语言引导的指称表达理解推理网络系统及推理方法

(57)摘要

本发明提供一种基于语言引导的指称表达理解推理网络系统及推理方法,包括:文本特征提取器、图像特征提取器、文本特征扩展器(TFE)、跨模态对齐模块(TCA)和跨模态融合模块(TCF);通过语言引导推理网络模型(LGR-NET),以充分利用指称表达式的指导;设置预测标记来捕捉跨模态特征,为了充分利用文本特征,通过文本特征扩展模块(TFE)从三个方面对其进行了扩展,文本生成的坐标嵌入有助于预测词元捕获关键的视觉特征;文本特征用于交替的跨模态推理;新颖的跨模态损失增强了跨模态对齐;如此文本特征从多个角度充分的引导了模型整体的跨模态推理流程,充分利用了文本中的线索,大大提高了模型性能。



CN 117609536 A

1.基于语言引导的指称表达理解推理网络系统,其特征在于,包括:文本特征提取器、图像特征提取器、文本特征扩展器、文本引导的跨模态对齐模块和文本引导的跨模态融合模块;

所述文本特征提取器:用于提取文本特征;

所述图像特征提取器:用于提取图像特征;

使用一个预测词元来捕获用于边界框预测的关键视觉和文本特征,并将其用于定位所指对象;为了充分捕捉指称表达中的线索,采用所述文本特征扩展器从三个方面扩展文本特征,即:生成坐标编码、词向量和句子向量;这三种文本特征随后通过被反复送入跨模态对齐模块和跨模态融合模块,参与跨模态对齐损失的计算来进行跨模态推理;扩展的三种文本特征在执行跨模态推理时被充分利用;

同时,预测词元也得到充分学习;预测头使用预测词元生成所指对象的边界框;

所述跨模态对齐模块:用于坐标嵌入、词嵌入和句子嵌入的输入和对齐损失的计算;

所述跨模态融合模块:用于坐标嵌入、词嵌入和句子嵌入三种文本特征的融合。

2.基于语言引导的指称表达理解推理网络系统的推理方法,其特征在于,包括如下步骤:

步骤一、多模态特征提取;

采用Swin Transformer作为图像特征提取器:

输入一张RGB图片$I \in \mathbb{R}^{3\times H\times W}$;其中3表示RGB通道数,H和W分别是图像的高和宽;

首先,通过图像块分割模块得到初始图像特征图$F_0 \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$;其中C为初始图像特征的通道数;

其次,通过四个降采样阶段:4倍、8倍、16倍、32倍;分别生成四个特征图[F_1,F_2,F_3,F_4];

然后,使用卷积核大小为1×1的卷积神经网络统一四个特征图的通道为D_v;通过在相邻的特征图上逐步使用mean pooling操作和取平均的操作最终得到一个聚合所有特征图的统一特征图$F_4^* \in \mathbb{R}^{D_v\times\frac{H}{32}\times\frac{W}{32}}$;

同时,为了提升大目标物体的特征提取效果,在特征图F_4^*的基础上通过2×2的max-pooling操作生成特征图$F_5^* \in \mathbb{R}^{D_v\times\frac{H}{64}\times\frac{W}{64}}$;

最后,将所述两个特征图F_4^*和F_5^*进行一维展平和拼接,得到图像特征$F_{v'} \in \mathbb{R}^{D_v\times N_v}$,其中特征长度$N_v = \frac{H}{32}\times\frac{H}{32}+\frac{H}{64}\times\frac{H}{64}$;

采用BERT作为文本特征提取器:

在经过词表映射后的文本向量前后分别添加了[CLS]和[SEP]词元;设置最大句子长度为N_t,经过BERT的特征提取,得到文本特征$F_{t'} \in \mathbb{R}^{D_t\times N_t}$;

在提取完图文特征后,分别使用两个全连接神经网络(FFN)将图文特征投射到同一个特征空间D,最终得到映射图像特征$F_v \in \mathbb{R}^{D\times N_v}$和映射文本特征$F_t \in \mathbb{R}^{D\times N_t}$,用于后续的模块输入;

步骤二、文本特征扩展;

为了充分利用文本特征进行跨模态推理,采用TFE模块来扩展文本特征;

TFE生成了空间特征的坐标编码,包含全体文本特征的词向量,以及浓缩特征的句子向量;前两者分别用于TCA和TCF模块,最后一个用于跨模态损失计算;

首先,生成一个新颖的坐标编码,以增强预测词元的空间表示;

使用[CLS]词元通过多层感知机(MLP)生成一个二维坐标;该过程的公式如下,

$$\begin{cases} coord = sigmoid(FFN_C(h_{cls})) \\ p_{coord} = PE(coord) \end{cases}$$

FFN包含两层线性层和一个ReLU激活函数;h_{cls}是[CLS]词元的特征表示,coord是归一化的二维坐标,PE函数将二维坐标转化为一个D维的正弦位置编码;

然后,TFE直接输出文本特征F_t作为词向量用于后续的TCF模块;

最后,TFE生成一个句子向量用于跨模态损失计算,句子向量由[CLS]向量生成,公式如下:

$f_{sent}=FFN_S(h_{cls})$;

步骤三、文本引导的跨模态对齐;

在之前提取的图像特征F_v前插入一个预测词元$f_p \in \mathbb{R}^{D\times 1}$作为初始的跨模态表示$X^0 \in \mathbb{R}^{D\times(1+N_v)}$,即:

$$X^0 = [f_p^0, \underbrace{f_1^0, f_2^0, \cdots, f_{N_v}^0}_{visual\ features}];$$

为了对齐图文特征,采用注意力机制;

在采用注意力机制前,需从指称表达中引入空间表示增强预测词元;即在每层的预测词元中加入了TFE中生成的坐标向量p_{coord};

$$\widehat{f_p^i} = f_p^i + p_{coord}$$

然后采用多头自注意力机制和残差链接、层归一化用于更新、对齐视觉特征;

$$\underbrace{Q_a = W_{Q_a}^T X^i, K_a = W_{K_a}^T X^i, V_a = W_{V_a}^T X^i}$$

$$\widetilde{X^i} = LN\left(softmax\left(\frac{Q_a K_a^T}{\sqrt{d_k}}\right)V_a + X^i\right)$$

其中:$W_{Q_a}, W_{K_a}, W_{V_a}$分别是query、key、value的映射权重矩阵;TCA的输出$\widetilde{X^i}$即为对齐的图像特征,d_k是通道数,LN(•)为层归一化;

步骤四、文本引导的跨模态融合;

采用跨模态注意力机制融合文本特征;

对齐的图像特征$\widetilde{X^i}$作为query,文本特征F_t作为key和value;文本引导的跨模态注意公式如下:

$$\underbrace{Q_f = W_{Q_f}^T \widetilde{X^i}, K_f = W_{K_f}^T F_t, V_f = W_{V_f}^T F_t}$$

$$\widetilde{X^i} = LN(softmax\left(\frac{Q_f K_f^T}{\sqrt{d_k}}\right)V_f + \widetilde{X^i})$$

$$X^{i+1} = LN(\overline{X^i} + FFN_F(\overline{X^i}))$$

其中:$W_{Q_f}, W_{K_f}, W_{V_f}$分别是query、key、value的映射权重矩阵;

此处,输出的跨模态表示X^{i+1}捕获了跟指称物体相关的关键文本特征;与此同时,预测词元聚合了视觉相关的文本特征用于后续的TCA;通过堆叠N层TCA和TCF模块,使它们交替工作;

步骤五、预测头;

基于最后一层输出的,经过充分学习的预测词元f_p^N,采用一个三层的FFN和sigmoid激活函数生成最终的预测框:

$$(x, y, w, h) = sigmoid(FFN_H(f_p^N));$$

其中,x,y分别代表预测框中心点的坐标,w,h表示预测框的宽和高;

步骤六、损失和训练;

为了训练LGR-NET,采用一个包含两项的损失函数;前者为预测框回归损失,后者为跨模态对齐损失,

$$\mathcal{L}_{total} = \sum_{i=1}^{N} \mathcal{L}_{box_i} + \lambda \sum_{i=1}^{N} \mathcal{L}_{align_i}$$

前者的损失用于帮助预测词元捕获所指对象的边缘特征;后者的损失促进其捕获与指称表达对象的语义一致性特征;超参数λ用于平衡两者;

为了增强指称目标和指称表达之间的对齐效果,采用对比形式的对齐损失;将一个批次里匹配的图文对视作正样本,不匹配的视作负样本;

对齐损失如下,

$$\mathcal{L}_{align} = -\frac{1}{B}\sum_{j=1}^{B} log\frac{\exp(f_{obj}^j \cdot f_{sent}^j/\tau)}{\sum_{k=1}^{B}\exp(f_{obj}^j \cdot f_{sent}^k/\tau)}$$

其中:B为批量大小,·代表内积;其中τ是可训练的温度参数,用于控制分布的平滑度;其中句子特征f_{sent}来自TFE的输出,物体特征来自预测词元f_p;

$f_{obj}=FFN_{ojb}(f_p)$

为了简化公式表达,层数下标被忽略。

3.根据权利要求2所述的基于语言引导的指称表达理解推理网络系统的推理方法,其特征在于,指称表达包括:位置词或对象之间的空间关系,使用空间信息可以有效地增强用于定位的预测词元的空间表示。

4.根据权利要求3所述的基于语言引导的指称表达理解推理网络系统的推理方法,其特征在于,本发明模型包括:N层堆叠的TCA和TCF模块,将第i层的预测框表示为$b_i=(x_i,y_i,w_i,h_i)$;真实的标签为$\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$,因此损失的第一项计算如下,

$$\mathcal{L}_{box_i} = \mathcal{L}_{GIoU}(b_i, \hat{b}) + \mathcal{L}_{L_1}(b_i, \hat{b});$$

其中,$\mathcal{L}_{GIoU}$和$\mathcal{L}_{L_1}$分别是GIoU损失和L1损失。

5.根据权利要求3所述的基于语言引导的指称表达理解推理网络系统的推理方法,其特征在于,使用预测准确率来评估推理网络模型的结果;即,当预测框和真实框的交并比(IoU)大于0.5时被认为是一次正确的预测,依此计算模型在测试集和验证集上的准确率。

基于语言引导的指称表达理解推理网络系统及推理方法

技术领域

[0001] 本发明属于计算机智能处理技术领域,尤其是一种基于语言引导的指称表达理解推理网络系统及推理方法。

背景技术

[0002] 指称表达理解(REC)是视觉和语言领域的一项基本任务,旨在根据自然语言表达定位图像区域,REC要求模型捕捉文本中的关键线索,并执行准确的跨模态推理。

[0003] 为了解决REC任务,一个关键挑战是如何执行准确的跨模态推理:现有技术中,通常采用如下三种方法进行跨模态推理,即:两阶段方法、一阶段方法和基于Transformer的REC方法;

[0004] ①两阶段方法:首先从图像中生成一组区域建议,然后使用跨模态相似性度量来测量候选区域与指称表达之间的匹配分数;最后选择具有最高匹配得分的区域作为最终的预测结果。

[0005] ②一阶段方法:通常在提取图像特征的同时执行多模态融合,并直接在预定义的锚点上预测边界框。

[0006] 总的来说,上述两种类型的方法都严重依赖于现成的对象检测器的性能,具体而言,前两类方法分别基于两阶段或一阶段对象检测器;两阶段流程通常首先生成图片的一组区域建议,然后通过检索与给定表达式的匹配得分最高的区域作为最终结果;相反,一阶段方法在提取图像特征上直接预测具有最大置信度得分的预定义锚点作为结果;这两种方法都基于通用的对象检测器,并在预先生成的候选对象上预测结果,因此,它们的性能通常受到生成的提议或预定义的锚点的限制。

[0007] ③基于Transformer的REC方法:

[0008] Transformer最初是为机器翻译提出的,并广泛用于各种自然语言处理任务;近期Transformer已经扩展到计算机视觉任务,例如图像分类和目标检测;基于Transformer的REC方法使用Transformer架构进行特征提取和多模态交互,并直接生成定位框;开创性的工作TransVG通过CNN骨干网络和Transformer编码器对图像特征进行编码。BERT被用于提取语言特征,并构建了一个Transformer编码器(称为视觉-语言Transformer)来融合拼接的图像-文本特征;这些框架采用堆叠的Transformer层进行跨模态推理,并直接回归边界框,而无需现成的检测器;与前面讲述的两阶段和一阶段方法相比,这些基于Transformer的方法更加优雅且性能更好。

[0009] 但大多数基于转换器的方法通常平等地对待图像和文本,它们通常以粗糙的方式进行跨模态推理,在没有详细考虑(例如空间信息)的情况下整体利用文本特征;这种对文本特征的不充分利用将导致次优结果。

[0010] 综上所述,在执行跨模态推理时,现有技术的方法通常将图像和文本特征等同对待,它们通常以简单的方式使用图像和文本特征,通过拼接方式进行同质化的推理;此外,文本特征被作为一个整体使用的,没有具体的区分;然而,我们认为图像和文本特征在REC

任务中发挥着不同的作用;指称表达是跨模态推理的重要指导,而图像是定位目标的载体;因此,REC模型需要捕捉文本中的重要线索,并需要使用这些线索结合图像逐渐进行推理,识别目标对象,最终在图像中定位它。

发明内容

[0011] 为了解决上述技术问题,本发明提供一种基于语言引导的指称表达理解推理网络系统及推理方法,提出了一种语言引导的推理网络模型(LGR-NET),以充分利用文本特征进行有效的跨模态推理,准确定位被引用的对象。

[0012] 基于语言引导的指称表达理解推理网络系统及推理方法,其中:

[0013] 基于语言引导的指称表达理解推理网络系统,包括:文本特征提取器、图像特征提取器、文本特征扩展器(TFE)、文本引导的跨模态对齐模块(TCA)和文本引导的跨模态融合模块(TCF);

[0014] 所述文本特征提取器:用于提取文本特征;

[0015] 所述图像特征提取器:用于提取图像特征;

[0016] 使用一个预测词元来捕获用于边界框预测的关键视觉和文本特征,并将其用于定位所指对象;为了充分捕捉指称表达中的线索,采用所述文本特征扩展器从三个方面扩展文本特征,即:生成坐标嵌入(坐标编码)、词嵌入(词向量)和句子嵌入(句子向量);这三种文本特征随后通过被反复送入跨模态对齐模块和跨模态融合模块,参与跨模态对齐损失的计算来进行跨模态推理;扩展的三种文本特征在执行跨模态推理时被充分利用;

[0017] 同时,预测词元也得到充分学习;预测头使用预测词元生成所指对象的边界框;

[0018] 所述跨模态对齐模块:用于坐标嵌入、词嵌入和句子嵌入的输入和对齐损失的计算;

[0019] 所述跨模态融合模块:用于坐标嵌入、词嵌入和句子嵌入三文本特征的融合;

[0020] 基于语言引导的指称表达理解推理网络系统的推理方法,包括如下步骤:

[0021] 步骤一、多模态特征提取;

[0022] 采用Swin Transformer作为图像特征提取器:

[0023] 输入一张RGB图片$I \in \mathbb{R}^{3\times H\times W}$;其中3表示RGB通道数,H和W分别是图像的高和宽;

[0024] 首先,通过图像块分割模块得到初始图像特征图$F_0 \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$;其中C为初始图像特征的通道数;

[0025] 其次,通过四个降采样阶段(4倍、8倍、16倍、32倍)分别生成四个特征图[F_1,F_2,F_3,F_4];

[0026] 然后,使用卷积核大小为1×1的卷积神经网络统一四个特征图的通道为D_v;通过在相邻的特征图上逐步使用mean pooling操作和取平均的操作最终得到一个聚合所有特征图的统一特征图$F_4^* \in \mathbb{R}^{D_v\times\frac{H}{32}\times\frac{W}{32}}$;

[0027] 同时,为了提升大目标物体的特征提取效果,在特征图F_4^*的基础上通过2×2的

max-pooling操作生成特征图$F_5^* \in \mathbb{R}^{D_v \times \frac{H}{64} \times \frac{W}{64}}$。

[0028] 最后，将所述两个特征图F_4^*和F_5^*进行一维展平和拼接，得到图像特征$F_{v'} \in \mathbb{R}^{D_v \times N_v}$，其中特征长度$N_v = \frac{H}{32} \times \frac{H}{32} + \frac{H}{64} \times \frac{H}{64}$。

[0029] 采用BERT作为文本特征提取器：

[0030] 在经过词表映射后的文本向量前后分别添加了[CLS]和[SEP]词元；设置最大句子长度为N_t，经过BERT的特征提取，得到文本特征$F_{t'} \in \mathbb{R}^{D_t \times N_t}$；

[0031] 在提取完图文特征后，分别使用两个全连接神经网络(FFN)将图文特征投射到同一个特征空间D，最终得到映射图像特征$F_v \in \mathbb{R}^{D \times N_v}$和映射文本特征$F_t \in \mathbb{R}^{D \times N_t}$，用于后续的模块输入；

[0032] 步骤二、文本特征扩展；

[0033] 为了充分利用文本特征进行跨模态推理，采用TFE模块来扩展文本特征；

[0034] TFE生成了空间特征的坐标编码，包含全体文本特征的词向量，以及浓缩特征的句子向量；前两者分别用于TCA和TCF模块，最后一个用于跨模态损失计算；

[0035] 首先，生成一个新颖的坐标编码，以增强预测词元的空间表示；

[0036] 作为一种举例说明，指称表达通常包括位置词或对象之间的空间关系，比如“前面的”或“5点钟位置”；使用这些空间信息可以有效地增强用于定位的预测词元的空间表示。

[0037] 使用[CLS]词元通过多层感知机(MLP)生成一个二维坐标；该过程的公式如下，

[0038] $$\begin{cases} coord = sigmoid(FFN_C(h_{cls})) \\ p_{coord} = PE(coord) \end{cases}$$

[0039] FFN包含两层线性层和一个ReLU激活函数；h_{cls}是[CLS]词元的特征表示，coord是归一化的二维坐标，PE函数将二维坐标转化为一个D维的正弦位置编码；

[0040] 然后，TFE直接输出文本特征F_t作为词向量用于后续的TCF模块；

[0041] 最后，TFE生成一个句子向量用于跨模态损失计算，句子向量由[CLS]向量生成，公式如下：

[0042] $f_{sent} = FFN_S(h_{cls})$；

[0043] 步骤三、文本引导的跨模态对齐；

[0044] 在之前提取的图像特征F_v前插入一个预测词元$f_p \in \mathbb{R}^{D \times 1}$作为初始的跨模态表示$X^0 \in \mathbb{R}^{D \times (1+N_v)}$，即：

[0045] $$X^0 = [f_p^0, \underbrace{f_1^0, f_2^0, \cdots, f_{N_v}^0}_{visual\ features}];$$

[0046] 为了对齐图文特征，采用注意力机制；

[0047] 进一步的，在采用注意力机制前，需从指称表达中引入空间表示增强预测词元；即在每层的预测词元中加入了TFE中生成的坐标向量p_{coord}；

[0048] $$\widehat{f_p^i} = f_p^i + p_{coord}$$

[0049] 然后采用多头自注意力机制和残差链接、层归一化用于更新、对齐视觉特征；

[0050] $$\underline{Q_a = W_{Q_a}^T X^i, K_a = W_{K_a}^T X^i, V_a = W_{V_a}^T X^i}$$

[0051] $$\widetilde{X^i} = LN(softmax\left(\frac{Q_a K_a^T}{\sqrt{d_k}}\right)V_a + X^i)$$

[0052] 其中：$W_{Q_a}, W_{K_a}, W_{V_a}$分别是query、key、value的映射权重矩阵；TCA的输出$\widetilde{X^i}$即为对齐的图像特征，d_k是通道数，LN(•)为层归一化；

[0053] 步骤四、文本引导的跨模态融合；

[0054] 采用跨模态注意力机制融合文本特征；

[0055] 对齐的图像特征$\widetilde{X^i}$作为query，文本特征F_t作为key和value；文本引导的跨模态注意公式如下：

[0056] $$\underline{Q_f = W_{Q_f}^T \widetilde{X^i}, K_f = W_{K_f}^T F_t, V_f = W_{V_f}^T F_t}$$

[0057] $$\widetilde{X^i} = LN(softmax\left(\frac{Q_f K_f^T}{\sqrt{d_k}}\right)V_f + \widetilde{X^i})$$

[0058] $$X^{i+1} = LN(\overline{X^i} + FFN_F(\overline{X^i}))$$

[0059] 其中：$W_{Q_f}, W_{K_f}, W_{V_f}$分别是query、key、value的映射权重矩阵；

[0060] 此处，输出的跨模态表示X^{i+1}捕获了跟指称物体相关的关键文本特征；与此同时，预测词元聚合了视觉相关的文本特征用于后续的TCA；通过堆叠N层TCA和TCF模块，使它们交替工作；

[0061] 步骤五、预测头；

[0062] 基于最后一层输出的，经过充分学习的预测词元f_p^N，采用一个三层的FFN和sigmoid激活函数生成最终的预测框：

[0063] $$(x, y, w, h) = sigmoid(FFN_H(f_p^N))$$；

[0064] 其中，x,y分别代表预测框中心点的坐标，w,h表示预测框的宽和高；

[0065] 步骤六、损失和训练；

[0066] 为了训练LGR-NET，采用一个包含两项的损失函数；前者为预测框回归损失，后者为跨模态对齐损失，

[0067] $$\mathcal{L}_{total} = \sum_{i=1}^{N} \mathcal{L}_{box_i} + \lambda \sum_{i=1}^{N} \mathcal{L}_{align_i}$$

[0068] 前者的损失用于帮助预测词元捕获所指对象的边缘特征；后者的损失促进其捕获与指称表达对象的语义一致性特征；超参数λ用于平衡两者；

[0069] 作为一种举例说明，我们的模型包括N层堆叠的TCA和TCF模块，将第i层的预测框表示为$b_i = (x_i, y_i, w_i, h_i)$；真实的标签为$\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$，因此损失的第一项计算如下，

[0070] $\mathcal{L}_{box_i} = \mathcal{L}_{GIoU}(b_i, \hat{b}) + \mathcal{L}_{L_1}(b_i, \hat{b})$;

[0071] 其中,$\mathcal{L}_{GIoU}$和$\mathcal{L}_{L_1}$分别是GIoU损失和L1损失。

[0072] 此外,为了增强指称目标和指称表达之间的对齐效果,采用对比形式的对齐损失;将一个批次里匹配的图文对视作正样本,不匹配的视作负样本;

[0073] 对齐损失如下,

[0074] $$\mathcal{L}_{align} = -\frac{1}{B}\sum_{j=1}^{B} log \frac{\exp(f_{obj}^{j} \cdot f_{sent}^{j}/\tau)}{\sum_{k=1}^{B}\exp(f_{obj}^{j} \cdot f_{sent}^{k}/\tau)}$$

[0075] 其中:B为批量大小,·代表内积。其中τ是可训练的温度参数,用于控制分布的平滑度。其中句子特征f_{sent}来自TFE的输出,物体特征来自预测词元f_p;

[0076] $f_{obj}=FFN_{ojb}(f_p)$

[0077] 为了简化公式表达,层数下标被忽略;

[0078] 作为一种举例说明,使用预测准确率来评估推理网络模型的结果;即,当预测框和真实框的交并比(IoU)大于0.5时被认为是一次正确的预测,依此计算模型在测试集和验证集上的准确率。

[0079] 本发明的有益效果:

[0080] 本发明提出了一个语言引导推理网络模型(LGR-NET),以充分利用指称表达式的指导;为了定位引用的对象,设置了一个预测词元来捕捉跨模态特征;此外,为了充分利用文本特征,通过文本特征扩展模块(TFE)从三个方面对其进行了扩展。

[0081] 提出了LGR-NET用于REC任务;LGR-NET强调从三个方面利用文本特征引导跨模态推理;文本生成的坐标嵌入有助于预测词元捕获关键的视觉特征;文本特征用于交替的跨模态推理;新颖的跨模态损失增强了跨模态对齐;如此文本特征从多个角度充分的引导了模型整体的跨模态推理流程,充分利用了文本中的线索,大大提高了模型性能。

附图说明

[0082] 图1为本发明基于语言引导的指称表达理解推理网络系统及推理方法之现有技术同质化推理方法和本发明推理方法在解决REC任务上的差异对比图。

[0083] 图2为本发明基于语言引导的指称表达理解推理网络系统的LGR-NET框架示意图。

[0084] 图3为本发明基于语言引导的指称表达理解推理网络系统的推理方法之流程总览图。

具体实施方式

[0085] 下面,参考图1至3图所示,基于语言引导的指称表达理解推理网络系统及推理方法,其中:

[0086] 基于语言引导的指称表达理解推理网络系统,包括:文本特征提取器、图像特征提取器、文本特征扩展器(TFE)、文本引导的跨模态对齐模块(TCA)和文本引导的跨模态融合模块(TCF);

[0087] 所述文本特征提取器:用于提取文本特征;

[0088] 所述图像特征提取器:用于提取图像特征;

[0089] 使用一个预测词元来捕获用于边界框预测的关键视觉和文本特征,并将其用于定位所指对象;为了充分捕捉指称表达中的线索,采用所述文本特征扩展器从三个方面扩展文本特征,即:生成坐标嵌入(坐标编码)、词嵌入(词向量)和句子嵌入(句子向量);这三种文本特征随后通过被反复送入跨模态对齐模块和跨模态融合模块,参与跨模态对齐损失的计算来进行跨模态推理;扩展的三种文本特征在执行跨模态推理时被充分利用;

[0090] 同时,预测词元也得到充分学习;预测头使用预测词元生成所指对象的边界框;

[0091] 所述跨模态对齐模块:用于坐标嵌入、词嵌入和句子嵌入的输入和对齐损失的计算;

[0092] 所述跨模态融合模块:用于坐标嵌入、词嵌入和句子嵌入三种文本特征的融合;

[0093] 基于语言引导的指称表达理解推理网络系统的推理方法,包括如下步骤:

[0094] 步骤一、多模态特征提取;

[0095] 采用Swin Transformer作为图像特征提取器:

[0096] 输入一张RGB图片$I \in \mathbb{R}^{3\times H\times W}$;其中3表示RGB通道数,H和W分别是图像的高和宽;

[0097] 首先,通过图像块分割模块得到初始图像特征图$F_0 \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$;其中C为初始图像特征的通道数;

[0098] 其次,通过四个降采样阶段(4倍、8倍、16倍、32倍)分别生成四个特征图[F_1,F_2,F_3,F_4];

[0099] 然后,使用卷积核大小为1×1的卷积神经网络统一四个特征图的通道为D_v;通过在相邻的特征图上逐步使用mean pooling操作和取平均的操作最终得到一个聚合所有特征图的统一特征图$F_4^* \in \mathbb{R}^{D_v\times\frac{H}{32}\times\frac{W}{32}}$;

[0100] 同时,为了提升大目标物体的特征提取效果,在特征图F_4^*的基础上通过2×2的max-pooling操作生成特征图$F_5^* \in \mathbb{R}^{D_v\times\frac{H}{64}\times\frac{W}{64}}$。

[0101] 最后,将所述两个特征图F_4^*和F_5^*进行一维展平和拼接,得到图像特征$F_{v'} \in \mathbb{R}^{D_v\times N_v}$,其中特征长度$N_v = \frac{H}{32}\times\frac{H}{32}+\frac{H}{64}\times\frac{H}{64}$。

[0102] 采用BERT作为文本特征提取器:

[0103] 在经过词表映射后的文本向量前后分别添加了[CLS]和[SEP]词元;设置最大句子长度为N_t,经过BERT的特征提取,得到文本特征$F_{t'} \in \mathbb{R}^{D_t\times N_t}$;

[0104] 在提取完图文特征后,分别使用两个全连接神经网络(FFN)将图文特征投射到同一个特征空间D,最终得到映射图像特征$F_v \in \mathbb{R}^{D\times N_v}$和映射文本特征$F_t \in \mathbb{R}^{D\times N_t}$,用于后续的模块输入;

[0105] 步骤二、文本特征扩展;

[0106] 为了充分利用文本特征进行跨模态推理,采用TFE模块来扩展文本特征;

[0107] TFE生成了空间特征的坐标编码,包含全体文本特征的词向量,以及浓缩特征的句子向量;前两者分别用于TCA和TCF模块,最后一个用于跨模态损失计算;

[0108] 首先,生成一个新颖的坐标编码,以增强预测词元的空间表示;

[0109] 作为一种举例说明,指称表达通常包括位置词或对象之间的空间关系,比如"前面的"或"5点钟位置";使用这些空间信息可以有效地增强用于定位的预测词元的空间表示。

[0110] 使用[CLS]词元通过多层感知机(MLP)生成一个二维坐标;该过程的公式如下,

[0111] $$\begin{cases} coord = sigmoid(FFN_C(h_{cls})) \\ p_{coord} = PE(coord) \end{cases}$$

[0112] FFN包含两层线性层和一个ReLU激活函数;h_{cls}是[CLS]词元的特征表示,coord是归一化的二维坐标,PE函数将二维坐标转化为一个D维的正弦位置编码;

[0113] 然后,TFE直接输出文本特征F_t作为词向量用于后续的TCF模块;

[0114] 最后,TFE生成一个句子向量用于跨模态损失计算,句子向量由[CLS]向量生成,公式如下:

[0115] $f_{sent}=FFN_S(h_{cls})$;

[0116] 步骤三、文本引导的跨模态对齐;

[0117] 在之前提取的图像特征F_v前插入一个预测词元$f_p \in \mathbb{R}^{D\times 1}$作为初始的跨模态表示$X^0 \in \mathbb{R}^{D\times(1+N_v)}$,即:

[0118] $$X^0 = [f_p^0, \underbrace{f_1^0, f_2^0, \cdots, f_{N_v}^0}_{visual\ features}];$$

[0119] 为了对齐图文特征,采用注意力机制;

[0120] 进一步的,在采用注意力机制前,需从指称表达中引入空间表示增强预测词元;即在每层的预测词元中加入了TFE中生成的坐标向量p_{coord};

[0121] $$\widehat{f_p^i} = f_p^i + p_{coord}$$

[0122] 然后采用多头自注意力机制和残差链接、层归一化用于更新、对齐视觉特征;

[0123] $$\underbrace{Q_a = W_{Q_a}^T X^i, K_a = W_{K_a}^T X^i, V_a = W_{V_a}^T X^i}$$

[0124] $$\widetilde{X^i} = LN\left(softmax\left(\frac{Q_a K_a^T}{\sqrt{d_k}}\right)V_a + X^i\right)$$

[0125] 其中:$W_{Q_a}, W_{K_a}, W_{V_a}$分别是query、key、value的映射权重矩阵;TCA的输出$\widetilde{X^i}$即为对齐的图像特征,d_k是通道数,LN(•)为层归一化;

[0126] 步骤四、文本引导的跨模态融合;

[0127] 采用跨模态注意力机制融合文本特征;

[0128] 对齐的图像特征$\widetilde{X^i}$作为query,文本特征F_t作为key和value;文本引导的跨模态注意公式如下:

[0129] $$\underbrace{Q_f = W_{Q_f}^T \widetilde{X^i}, K_f = W_{K_f}^T F_t, V_f = W_{V_f}^T F_t}$$

[0130] $$\widetilde{X^i} = LN(softmax\left(\frac{Q_f K_f^T}{\sqrt{d_k}}\right)V_f + \widetilde{X^i})$$

[0131] $$X^{i+1} = LN(\overline{X^i} + FFN_F(\overline{X^i}))$$

[0132] 其中:$W_{Q_f}, W_{K_f}, W_{V_f}$分别是query、key、value的映射权重矩阵;

[0133] 此处,输出的跨模态表示X^{i+1}捕获了跟指称物体相关的关键文本特征;与此同时,预测词元聚合了视觉相关的文本特征用于后续的TCA;通过堆叠N层TCA和TCF模块,使它们交替工作;

[0134] 步骤五、预测头;

[0135] 基于最后一层输出的,经过充分学习的预测词元f_p^N,采用一个三层的FFN和sigmoid激活函数生成最终的预测框:

[0136] $$(x, y, w, h) = sigmoid(FFN_H(f_p^N));$$

[0137] 其中,x,y分别代表预测框中心点的坐标,w,h表示预测框的宽和高;

[0138] 步骤六、损失和训练;

[0139] 为了训练LGR-NET,采用一个包含两项的损失函数;前者为预测框回归损失,后者为跨模态对齐损失,

[0140] $$\mathcal{L}_{total} = \sum_{i=1}^{N}\mathcal{L}_{box_i} + \lambda\sum_{i=1}^{N}\mathcal{L}_{align_i}$$

[0141] 前者的损失用于帮助预测词元捕获所指对象的边缘特征;后者的损失促进其捕获与指称表达对象的语义一致性特征;超参数λ用于平衡两者;

[0142] 作为一种举例说明,我们的模型包括N层堆叠的TCA和TCF模块,将第i层的预测框表示为$b_i = (x_i, y_i, w_i, h_i)$;真实的标签为$\hat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h})$,因此损失的第一项计算如下,

[0143] $$\mathcal{L}_{box_i} = \mathcal{L}_{GIoU}(b_i, \hat{b}) + \mathcal{L}_{L_1}(b_i, \hat{b});$$

[0144] 其中,$\mathcal{L}_{GIoU}$和$\mathcal{L}_{L_1}$分别是GIoU损失和L1损失。

[0145] 此外,为了增强指称目标和指称表达之间的对齐效果,采用对比形式的对齐损失;将一个批次里匹配的图文对视作正样本,不匹配的视作负样本;

[0146] 对齐损失如下,

[0147] $$\mathcal{L}_{align} = -\frac{1}{B}\sum_{j=1}^{B} log\frac{\exp(f_{obj}^j \cdot f_{sent}^j/\tau)}{\sum_{k=1}^{B}\exp(f_{obj}^j \cdot f_{sent}^k/\tau)}$$

[0148] 其中:B为批量大小,·代表内积。其中τ是可训练的温度参数,用于控制分布的平滑度。其中句子特征f_{sent}来自TFE的输出,物体特征来自预测词元f_p;

[0149] $f_{obj} = FFN_{ojb}(f_p)$

[0150] 为了简化公式表达,层数下标被忽略;

[0151] 作为一种举例说明,使用预测准确率来评估推理网络模型的结果;即,当预测框和

真实框的交并比(IoU)大于0.5时被认为是一次正确的预测,依此计算模型在测试集和验证集上的准确率。

[0152] 本发明提出了一个语言引导推理网络模型(LGR-NET),以充分利用指称表达式的指导;为了定位引用的对象,设置了一个预测标记来捕捉跨模态特征;此外,为了充分利用文本特征,通过我们的文本特征扩展模块(TFE)从三个方面对其进行了扩展。

[0153] 提出了LGR-NET用于REC任务;LGR-NET强调从三个方面利用文本特征引导跨模态推理;文本生成的坐标嵌入有助于预测词元捕获关键的视觉特征;文本特征用于交替的跨模态推理;新颖的跨模态损失增强了跨模态对齐;如此文本特征从多个角度充分的引导了模型整体的跨模态推理流程,充分利用了文本中的线索,大大提高了模型性能。

[0154] 为了更好的理解本发明的原理,现通过具体实施例举例说明如下,参照图1所示:

[0155] 例如:本发明简明地展示了两种REC框架,根据一个摆满橘子的图片和相应的指称表达生成边界框;作为一个基本的视觉-语言任务,REC可以推动各种应用,包括图像描述、视觉问题回答(VQA)和视觉导航。

[0156] 以图1为例,模型需要捕获关键线索,包括“橘子”和“5点钟位置”;前者指的是对象,后者指示空间信息;

[0157] 在关键线索的指导下,模型可以“理解”要定位的是什么以及哪一个;因此,使用现有技术的同质化推理方案是不够的,容易产生推理偏差,如图1所示现有技术图像定位错误,定位了左下角的橘子,即五点钟方向的语言文本信息推理为七点钟方向,尤其是当参照表达很复杂时,推理容易产生错误;

[0158] 因此,本发明采用充分利用文本特征进行准确的跨模态推理,使得推理结果准确无误。

[0159] 以上所述的仅为本发明的优选实施例,所应理解的是,以上实施例的说明只是用于帮助理解本发明的方法及其核心思想,并不用于限定本发明的保护范围,凡在本发明的思想和原则之内所做的任何修改、等同替换等等,均应包含在本发明的保护范围之内。

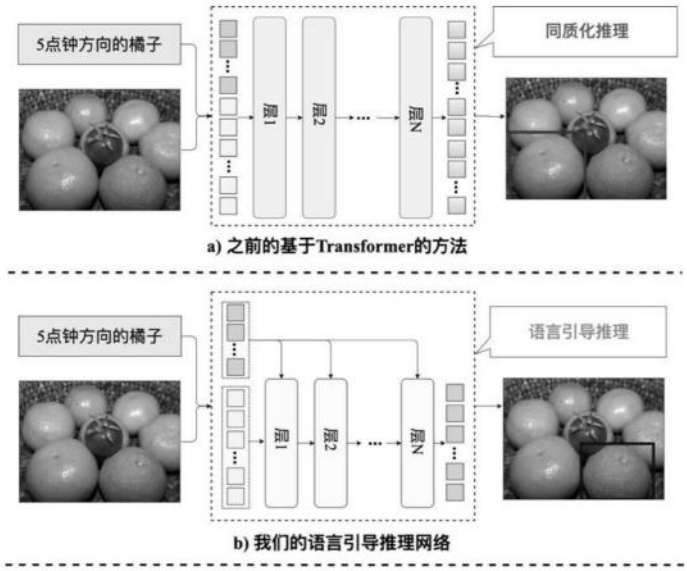

5点钟方向的橘子
层1
层2
层N
同质化推理
a) 之前的基于Transformer的方法
5点钟方向的橘子
层1
层2
层N
语言引导推理
b) 我们的语言引导推理网络

图1

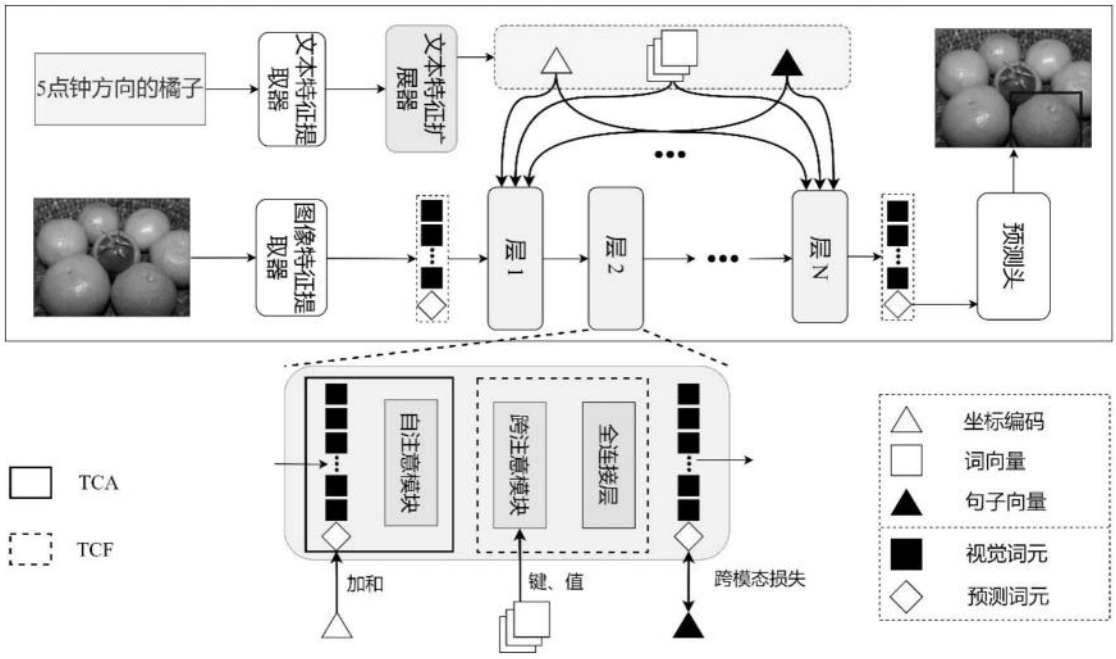

5点钟方向的橘子
文本特征提取器
文本特征扩展器
图像特征提取器
层 1
层 2
层 N
预测头
TCA
TCF
自注意模块
跨注意模块
全连接层
加和
键、值
跨模态损失
坐标编码
词向量
句子向量
视觉词元
预测词元

图2

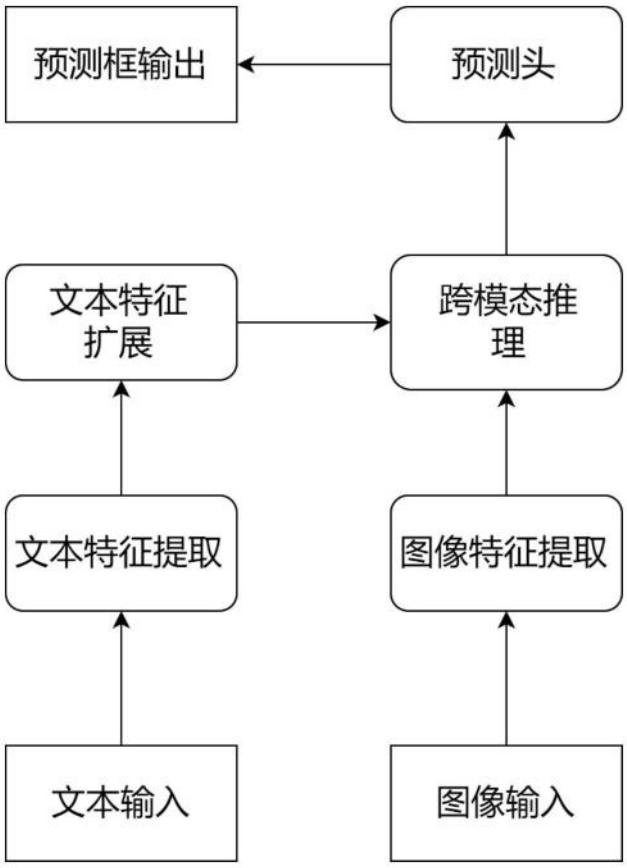


图3