计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220019
60(11): 2611−2623, 2023

一种面向指代短语理解的关系聚合网络

郭文雅　张　莹　刘胜哲　杨巨峰　袁晓洁
（南开大学计算机学院　天津　300350）
（guowenya@dbis.nankai.edu.cn）

Relationship Aggregation Network for Referring Expression Comprehension

Guo Wenya, Zhang Ying, Liu Shengzhe, Yang Jufeng, and Yuan Xiaojie
（*College of Computer Science*, *Nankai University*, *Tianjin* 300350）

Abstract　In this paper, we focus on the task of referring expression comprehension (REC), which aims to locate the corresponding regions in images referred by expressions. One of the main challenges is to visually ground the object relationships described by the input expressions. The existing mainstream methods mainly score objects based on their visual attributes and the relationships with other objects, and the object with the highest score is predicted as the referred region. However, these methods tend to only consider the relationships between the current evaluated region and its surroundings, but ignore the informative interactions among the multiple surrounding regions, which are important for matching the input expressions and visual content in image. To address this issue, we propose a relationship aggregation network (RAN) to construct comprehensive relationships and then aggregate them to predict the referred region. Specifically, we construct both the two kinds of aforementioned relationships based on graph attention networks. Then, the relationships most relevant to the input expression are selected and aggregated with a cross-modality attention mechanism. Finally, we compute the matching scores according to the aggregated features, based on which we predict the referred regions. Additionally, we improve the existing erase strategies in REC by erasing some continuous words to encourage the model find and use more clues. Extensive experiments on three widely-used benchmark datasets demonstrate the superiority of the proposed method.

Key words　referring expression comprehension; attention mechanism; graph attention network; modular network; erasing strategy

摘　要　指代短语理解（referring expression comprehension, REC）任务的目的是定位输入短语所指代的图像区域，其中最主要的挑战之一是在图像中建立和定位由输入短语描述的物体之间的关系. 现有的主流方法之一是根据物体本身的特性以及与其他物体的关系对当前物体进行打分，将得分最高的物体作为预测的被指代区域. 然而，这类方法往往只考虑物体与其周围环境之间的关系，而忽略了输入短语中所描述的周围环境之间的交互关系，这大大影响了对物体间关系的建模. 为了解决这一问题，提出了关系聚合网络（relationship aggregation network, RAN）来构建物体之间的关系，进而预测输入短语所指代的内容. 具体来说，利用图注意力网络建模图像物体之间完备的关系；然后利用跨模态注意力方法选择与输入短语最相关的关系进行聚合；最后，计算目标区域与输入短语之间的匹配分数. 除此之外，对指代短语理解中的擦除方法进行了改进，通过自适应扩充擦除范围的方式促使模型利用更多的线索来定位正确的区域. 在3个广泛使用的基准数据集上进行了大量的实验，结果证明了所提出方法的优越性.

收稿日期：2022-01-04；**修回日期**：2023-01-09
基金项目：国家自然科学基金-联合基金（U1903128）
This work was supported by the National Natural Science Foundation of China-Joint Fund (U1903128).
通信作者：张莹（yingzhang@nankai.edu.cn）

关键词 指代短语理解;注意力机制;图注意模型;模块化网络;擦除策略

中图法分类号 TP391

指代短语理解(referring expression comprehension, REC)要求在图像中定位短语指代的物体[1]. REC 可以被广泛应用于其他视觉理解任务,如视觉问答[2-4]、图像描述生成[5-7]. 指代短语理解的核心在于 2 个方面:1)找到具有短语中所述类别和属性的物体;2)根据短语中定义的物体之间的关系定位正确的区域. 尽管卷积神经网络,如 VGGNet[8] 和 ResNet[9],能够很好地识别图像物体的属性,但是精确定位短语所描述的关系仍然非常具有挑战性.

近年来,对关系建模的探索主要集中在 2 个方面:文献 [10–12] 使用图来表示图像中的关系,并据此定位短语所指的区域. 然而,将图像转换为图的过程中可能会丢失一些微妙但关键的信息,并且进一步确定被指代的区域仍然需要分析复杂的图结构. 在一些其他研究中[1,13-14],研究人员用一种更直接的方式来模拟图像中物体之间的关系,自适应地提取短语中对关系的描述,并将其与预先定义的视觉特征相关联. 在这类方法中,当前待评估物体(以下称为“候选物体”)所具备的关系被定义为与其周围物体(以下称为“上下文物体”)之间的相对位置. 而后,通过基于短语的跨模态注意力机制学习不同“上下文-候选关系”的权重,根据得到的权重对“上下文-候选关系”进行聚合,与短语中的文本描述计算匹配分数,以此评估当前候选物体是否为短语所指代的正确区域,如图 1 所示.

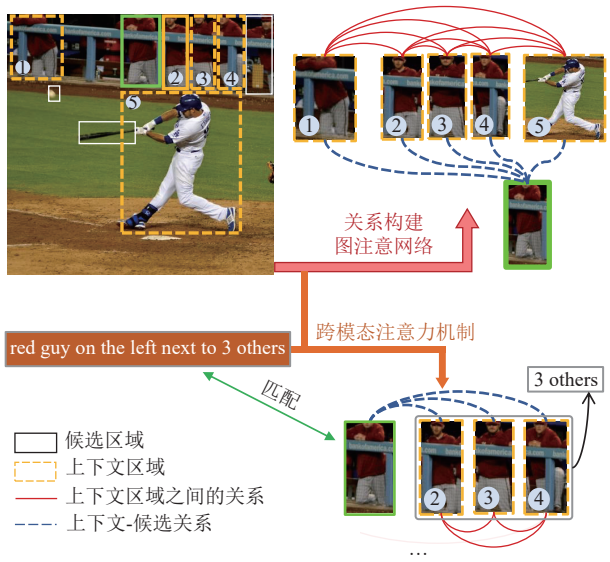


Fig. 1 Illustration of relationship aggregation in REC

图 1 指代短语理解中的关系聚合示意图

然而,由于上下文区域之间并非完全独立,在指代短语理解中,仅考虑上下文候选关系是不够的. 如图 1 所示,实线框内为当前候选物体,即:模型将要对该物体进行打分,编号为 1~5 的 5 个物体为当前候选物体周围的上下文物体. 要定位图像中的“red guy on the left next to 3 others”(左边的穿着红衣服挨着另外 3 个人的人),模型必须理解并建立 2 个关系:“left”(在左边)和“next to”(挨着). 虽然“在左边”的关系可以从候选物体与其上下文之间的关系推断出来,但是理解和定位到“穿红色衣服的人”具体“挨着”什么内容仍然非常具有挑战性. 而目标“3 others”(另外 3 个人)是包括编号为 2~4 的区域的整体. 当只考虑上下文-候选关系时,编号为 2~4 物体被分别单独处理,很难将它们识别为一个群体. 相比之下,由于编号为 2~4 的对象彼此接近,利用上下文物体之间的关系,分析它们之间的相对空间关系,很容易就能识别出“另外 3 个人”具体包括哪些区域. 综上所述,在指代短语理解的过程中,考虑上下文之间的相互关系,模型能够根据更加完整的信息准确地定位到短语指代的内容.

针对以上问题,本文提出了一个基于注意的关系聚合网络(relationship aggregation network, RAN)以便更好地建模图像物体之间的关系. 利用候选区域与上下文区域的位置关系来表示“上下文-候选”关系,并利用图注意网络(graph attention network, GAT)[15-16]来建模上下文区域之间的关系. 在此基础上,根据建立的关系与输入短语的相似度来对建模的关系进行聚合,进而根据聚合后的关系来衡量当前候选物体的得分.

此外,多项研究表明,注意力机制倾向于关注最重要的内容,容易忽略掉其他的补充信息[14,17],而输入的指代短语往往会从多个角度描述目标对象(例如,图 1 中的“left”和“next to 3 others”). 研究人员尝试用擦除策略来促使模型从除了最重要的角度外,挖掘更多的有利于找到正确物体的内容[14]. 但是本文发现,Liu 等人[14] 只擦除输入短语权重最高的一个单词,这种方法并不能很好地遮盖掉一个完整的描述物体的角度. 因此,本文对其擦除策略进行了改进,自适应地擦除多个单词以便完全擦除在一个角度上的描述. 利用最重要的单词来定位到该部分的位置,并在其周围寻找连续的单词来进行擦除操作.

综上所述，本文的贡献总结为 3 方面：

1）提出了一个基于注意的关系聚合网络，在指代短语理解任务中，同时考虑了“上下文-候选关系”和“上下文-上下文关系”.

2）设计了一种新的擦除策略，可以完整地擦除指代短语中在一个角度上的描述，促使模型利用更多的线索来定位正确的区域.

3）在 3 个常用数据集上进行的大量实验表明，本文提出的方法与最先进的方法相比具有良好的性能.

1 相关工作

本节主要介绍与所提出的方法密切相关的指代短语理解的最新方法.

1.1 指代短语理解

指代短语理解的目标是在图像中定位出指代短语所描述的区域[1]. 最近的一些工作[18-22]将此任务与指代短语生成任务相关联，并构建 CNN/LSTM 结构来处理图像区域和指代短语，不同的上下文信息被用来辅助对视觉信息的理解. 在文献 [20] 中，整个图像被用作上下文信息；而在文献 [21] 中，多实例学习获得的区域被用来建模图像区域之间的关联，类似地，对象之间的差异也用于表示视觉上下文. 在文献 [22] 中，研究人员构建了一个“表达”-“听取”的结构，利用 2 个任务之间的关联建模指代短语和图像区域的关系.

另外一些方法侧重于衡量输入短语和视觉区域特征在公共特征空间中的兼容性[18,23-24]. 但是这些方法忽略了图像和指代短语复杂的结构. 为了克服这一局限性，一些研究人员对表达式进行分解，并使用模块化网络来处理不同的组件. 这种模块化网络已成功应用于包括视觉问答[25-27]、视觉推理[28-29]和多任务强化学习[30]等多种任务中. 在指代短语理解任务中，早期的方法依赖于外部语言解析器[25-26,30]，后来一些工作通过注意机制进行语言分解[1,13,31]. Hu 等人[13]将表达式解析为主语、关系和宾语. 由于并非所有表达式都满足“主语、关系和宾语”的模板，后来的研究人员提出了更灵活的分解方案. 在文献 [1] 中，表达式被分解为主语、关系和位置. Wang 等人[31]进一步将图像中的关系分为同类物体之间的关系和不同类物体之间的关系. 除了模块化的处理方式，另一种对复杂数据建模的方法是使用图结构[10,12,32-33]或树结构[34]来表示图像中的区域和引用表达式中的单词之间的关系. 上述方法都是基于预先提取好的视觉区域，提取区域的质量影响模型效果，为了克服这一限制，最近开始有一些研究转向了单阶段的指代短语理解方法，直接从图像预测被指代的区域[35-37].

1.2 指代短语理解中的注意力机制

作为一种有效的深度学习技术，注意力机制常被用来提取与图像相关的文本内容和与语言描述相关的重要图像区域[38-40]. Deng 等人[38]提出了一种针对指代短语、图像内容和候选区域的累积注意机制；而在文献 [40] 中，对图像内容和视觉区域的注意是并行进行的. 最近，Hu 等人[39]利用双向跨模态注意模块更好地学习跨模态关系.

然而，这些方法主要关注图像中的视觉内容，而忽略了对象之间的交互作用. 为了更好地处理对象间的关系和提升模型对图像的理解能力，本文提出了一种基于注意的关系聚合网络. 图注意网络被用来构建一套完整的关系，从中可以提取关键信息内容，以提高学习与文本的匹配效果.

2 关系聚合网络

本文提出了一个基于注意的关系聚合网络来完成指代短语理解任务，通过充分考虑图像中对象之间的关系来定位输入短语指代的区域. 如图 2 所示，RAN 首先将输入图像表示为一系列物体区域，为当前候区域选择其周围的 5 个邻居作为上下文区域，利用 LSTM 提取输入表达式的文本特征，构建物体之间的关系；然后利用注意机制提取和聚集与短语相关的重要关系，聚合这些关系并计算表达式的相似性分数.

给定一个图像 I 包含 N 个图像区域，即 $I=\{R_i\}_{i=1}^{N}$，本文的目标是预测出由给定短语 E 所指代的区域 R^*. 本文为所有区域的计算与短语匹配分数，将分数最高的区域作为最终预测结果.

2.1 特征编码

在本节中，本文计算图像区域和输入指代短语的特征表示，在接下来的关系构建和关系聚合模块，模型为每个候选区域（可以是图像中的任何物体）计算与输入短语的匹配分数.

1）视觉特征抽取

为了更好地建模对象之间的关系，本文将每个图像 I 表示为一系列图像区域特征. 每个区域特征包括对应的视觉内容和在原始图像中的位置特征. 与之前做法[14]相同，本文使用以 ResNet101[41]为基础网络的 Faster R-CNN[9]来生成区域的视觉表示. 具体来

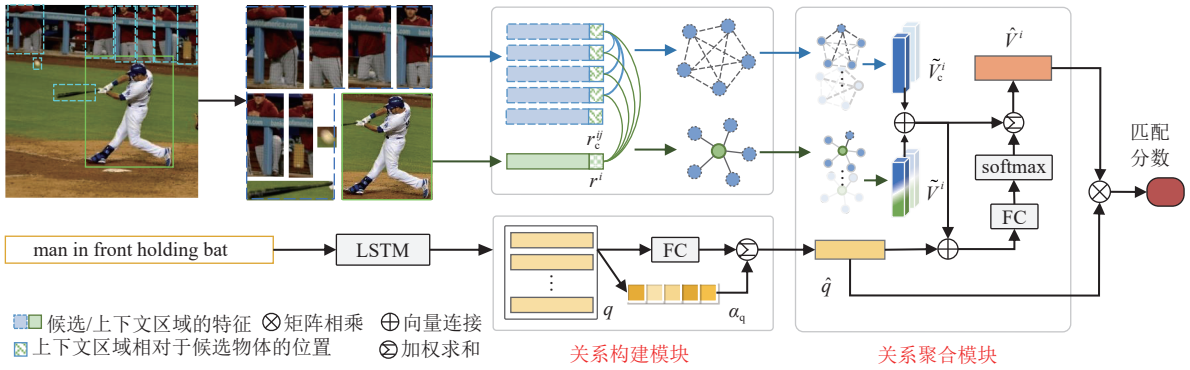


Fig. 2 Illustration of RAN

图 2 关系聚合网络示意图

说, 本文从 Faster R-CNN 的 C3 和 C4 层提取特征, C3 提取的特征包含低级颜色特征, 而 C4 包含高级类别特征, 最终区域 R_i 的视觉特征被表示为 $\boldsymbol{v}^i$. 为了进一步理解图像中的空间关系, 本文使用一个 5 维向量来表示区域在原始图像中的绝对位置. 区域的绝对位置表示为

$$\boldsymbol{l}^i=\left(\frac{x_{\mathrm{tl}}}{W},\frac{y_{\mathrm{tl}}}{H},\frac{x_{\mathrm{br}}}{W},\frac{y_{\mathrm{br}}}{H},\frac{w\cdot h}{W\cdot H}\right), \tag{1}$$

其中 $(x_{\mathrm{tl}}, y_{\mathrm{tl}})$ 和 $(x_{\mathrm{br}}, y_{\mathrm{br}})$ 分别用来指代区域左上角和右下角的坐标. 于是, 候选物体 R_i 可以表示为 $\boldsymbol{r}^i=(\boldsymbol{v}^i, \boldsymbol{l}^i)$.

本文用相对位置来表示区域之间的关系. 对于要评估的当前区域(即候选区域), 本文在它周围选择最多 5 个相邻区域(即上下文区域). 每个上下文区域的相对位置可以计算为

$$\delta\boldsymbol{l}^{ij}=\left(\frac{\Delta[x_{\mathrm{tl}}]_{ij}}{w_i},\frac{\Delta[y_{\mathrm{tl}}]_{ij}}{h_i},\frac{\Delta[x_{\mathrm{br}}]_{ij}}{w_i},\frac{\Delta[y_{\mathrm{br}}]_{ij}}{h_i},\frac{w_j\cdot h_j}{w_i\cdot h_i}\right), \tag{2}$$

那么 R_i 的第 j 个上下文区域的特征可以表示为 $\boldsymbol{r}_{\mathrm{c}}^{ij}=(\boldsymbol{v}^j, \delta\boldsymbol{l}^{ij})$. 以下关于关系的操作基于这些获得的视觉特征以及位置特征进行.

2) 短语特征抽取

对于具有 m 个单词的短语, 本文用预训练好的 Glove[42] 向量来表示每一个单词. 整个短语可以表示为一系列单词向量的集合: $Q=\{\boldsymbol{x}_i\}_{i=1}^m, \boldsymbol{x}_i\in\mathbb{R}^{d_{\mathrm{w}}}$. 没有在 Glove 词表中的单词则被随机初始化. 然后本文将单词的词向量输入到双向 LSTM 网络中, LSTM 隐含层特征维度为 d_{q}. 此外, 为了将短语的特征与图像相关内容进行关联, 本文用整个图像的特征 $\boldsymbol{v}_0=CNN(I)$, 来初始化 LSTM.将 LSTM 输出的向量作为短语的特征表示, $\boldsymbol{h}_t=(\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t})$.

2.2 关系构建(relationship construction, RC)模块

基于获得的特征, 在本节中 RAN 构造了完备的物体之间的关系, 以便更好地理解图像中的对象与输入短语之间的关联.

1) 视觉关系构建

物体之间的关系包括它们的类别属性以及它们之间的相互作用关系. 物体的属性信息由视觉特征来表示, 而相互作用关系则由相对位置来表示. 为了方便接下来的操作, 将 5 维的位置特征(包含候选物体的绝对位置和其上下文信息的相对位置)通过一个全连接层映射为 d_{v} 维, 得到 $\bar{\boldsymbol{l}}^i=fc(\boldsymbol{l}^i)$, $\delta\bar{\boldsymbol{l}}^{ij}=fc(\delta\boldsymbol{l}^{ij})$. 而后, 当前候选物体 R_i 所具有的"上下文-候选"关系可以表示为, $\boldsymbol{V}^i=(\bar{\boldsymbol{r}}^{ij}, \bar{\boldsymbol{r}}_{\mathrm{c}}^{ij})_{j=1}^5$, 其中 $\bar{\boldsymbol{r}}^{ij}=(\boldsymbol{v}^i, \bar{\boldsymbol{l}}^i)$, $\bar{\boldsymbol{r}}_{\mathrm{c}}^{ij}=(\boldsymbol{v}^j, \delta\bar{\boldsymbol{l}}^{ij})$. 类似地, 本文也可以获得上下文之间关系的特征表示 $\boldsymbol{V}_{\mathrm{c}}^i=(\bar{\boldsymbol{r}}_{\mathrm{c}}^{ij})_{j=1}^5$. 本文采用一个图注意层 $f_{\mathrm{GAT}}(\cdot)$ 来捕获上下文物体之间的关系. 图注意层可以利用输入的节点特征以及节点之间的邻接矩阵, 学习节点之间的关联, 输出一组新的与输入特征具有相同维度的特征, 该特征中包含节点之间的关联. 为建立上下文物体之间的关联, 本文将上下文区域的特征 $\boldsymbol{V}_{\mathrm{c}}^i$, 以及构建 5×5 的邻接矩阵 $\boldsymbol{M}_{\mathrm{c}i}$ 作为图注意层的输入($\boldsymbol{M}_{\mathrm{c}i}$ 被初始化为全 1 的邻接矩阵, 表示考虑每 2 个上下文物体之间的关系), 从而得到的所有上下文物体之前的关系为 $\tilde{\boldsymbol{V}}_{\mathrm{c}}^i=f_{\mathrm{GAT}}(\boldsymbol{V}_{\mathrm{c}}^i, \boldsymbol{M}_{\mathrm{c}i})$. $\tilde{\boldsymbol{V}}_{\mathrm{c}}^i$ 中每个元素 $\tilde{\boldsymbol{r}}_{\mathrm{c}}^{ij}$ 包含第 j 个上下文物体与其他上下文物体之间的关联.

2) 短语内关系构建

为了更好地建模指代短语中的语义关联, 本文采用了尺度点乘注意力层(scaled dot-product attention, SDPA)和前馈层神经网络层来计算 $\boldsymbol{q}$ 中每个元素的重要度, 包含查询 $\boldsymbol{Q}$、关键值 $\boldsymbol{K}$ 和特征值 $\boldsymbol{V}$. 尺度点乘注意力计算方法为

$$f_{\mathrm{SDPA}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{3}$$

其中 d_k 是输入的查询 $\boldsymbol{Q}$ 和关键值 $\boldsymbol{K}$. 前馈神经网络层(feed-forward network, FFN)计算方法为:

$$f_{\mathrm{FFN}}(x)=f_{\mathrm{relu}}(x\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2. \tag{4}$$

基于此，短语内单词之间的关系可以被计算为 $\tilde{\boldsymbol{q}}=f_{\mathrm{FFN}}(f_{\mathrm{SDPA}}(\boldsymbol{q},\boldsymbol{q},\boldsymbol{q}))$. 与从 LSTM 中输出的特征相比，$\tilde{\boldsymbol{q}}=(\tilde{\boldsymbol{h}}_t)_{t=1}^{m}$每一个单词的特征表示$\tilde{\boldsymbol{h}}_t$都包含了与其上下文更强的语义关联性.

2.3 关系聚合（relationship aggregation, RA）模块

尽管关系构建模块可以有效地模拟图像内部和文本中的关系，但是 REC 中还有另外一个关键点，即重要区域与指代短语之间的跨模态的关联. 如图 1 所示，这些关系对计算当前候选对象和表达式的匹配分数具有不同的贡献. 在本节中，本文将介绍如何根据指代短语抽取关键的关系信息.

首先，本文利用一个注意力层将之前获取的短语特征$\tilde{\boldsymbol{q}}$表示为一个单独的向量. 这个注意力层的计算方式为

$$\boldsymbol{\alpha}_{\mathrm{q}}=softmax(MLP(\tilde{\boldsymbol{q}})), \tag{5}$$

其中$\hat{\boldsymbol{q}}=\sum_{t=1}^{l}\boldsymbol{\alpha}_{\mathrm{q}_t}\tilde{\boldsymbol{h}}_t$, $\boldsymbol{\alpha}_{\mathrm{q}}=(\boldsymbol{\alpha}_{\mathrm{q}1},\boldsymbol{\alpha}_{\mathrm{q}2},...,\boldsymbol{\alpha}_{\mathrm{q}m})$, 为学习到的短语中每个单词的权重.

基于获得的$\hat{\boldsymbol{q}}$, 本文设计了跨模态注意力机制来选择并聚合重要的视觉关系：

$$\boldsymbol{\alpha}_{\mathrm{v}}^{i}=softmax(MLP([\hat{\boldsymbol{q}},\ \tilde{\boldsymbol{V}}_{\mathrm{c}}^{i},\boldsymbol{V}^{i}]), \tag{6}$$

$$\hat{\boldsymbol{V}}^{i}=\sum_{j=1}^{5}\boldsymbol{\alpha}_{\mathrm{v}j}^{i}[\bar{\boldsymbol{r}}_{\mathrm{c}}^{ij},\bar{\boldsymbol{r}}^{i}]. \tag{7}$$

得到的$\hat{\boldsymbol{V}}^{i}$被用来计算与$\hat{\boldsymbol{q}}$的匹配分数.

2.4 匹配分数计算

本节为每一个候选物体计算与输入短语的匹配分数. 理所当然地，如果当前的候选区域为短语所指代的区域，那么它将具有与短语的最高匹配分数.

遵循 Yu 等人[1]的做法，本文采用模块化的方式对区域的匹配分数进行计算. 除了物体本身的性质外，对于物体所具有的关系，人们通常会使用相对位置来区别与其同一类别的其他物体（如“man on the left”）. 因此，在计算匹配分数的过程中，使用 3 个模块来对当前候选物体进行评价：主体模块（以下缩写为 sub）、位置关系模块（以下缩写为 loc）和其他类型的关系（以下缩写为 rel）. 这 3 个模块的权重自适应地从输入短语中学习得来：$(w_{\mathrm{sub}},w_{\mathrm{loc}},w_{\mathrm{rel}})=softmax(f_{\mathrm{c}}(\boldsymbol{h}_m))$.

本文利用注意力机制将输入短语自适应地分解为 3 部分 $\boldsymbol{q}_{\mathrm{sub}}$, $\boldsymbol{q}_{\mathrm{loc}}$, $\boldsymbol{q}_{\mathrm{rel}}$, 分别用来计算不同模块中的匹配分数. 而不同模块中的视觉特征是区域的视觉特征以及位置特征衍生出来的. 主体模块用来处理区域的类别、属性等信息，不考虑物体之间的关联，所用的视觉特征为$\boldsymbol{V}_{\mathrm{sub}}^{i}=\boldsymbol{v}^{i}$; 位置关系经常被用来区分具有相同类别的物体，在位置关系模块中，本文着重考察物体之间的相对位置关系，使用的视觉特征包括构造的“上下文-候选”关系以及“上下文之间”的关系，即$\boldsymbol{V}_{\mathrm{loc}}^{i}=f_{\mathrm{RA}}(f_{\mathrm{RC}}(\boldsymbol{l}^{i},\delta\boldsymbol{l}^{i}))$并且 $\boldsymbol{V}_{\mathrm{rel}}^{i}=f_{\mathrm{RA}}(f_{\mathrm{RC}}(\boldsymbol{r}^{i},\boldsymbol{r}_{\mathrm{c}}^{i}))$, 其中，$f_{\mathrm{RC}}(\cdot)$和$f_{\mathrm{RA}}(\cdot)$分别代表关系构造模块和关系聚合模块. 对于不同的模块，本文利用函数$f_{\mathrm{m}}(a,b)=a\cdot b$来计算相应的视觉特征和文本内容之间的匹配分数：$S_{\mathrm{sub}}=f_{\mathrm{m}}(\boldsymbol{q}_{\mathrm{sub}},\boldsymbol{V}_{\mathrm{sub}}^{i})$, $S_{\mathrm{loc}}=f_{\mathrm{m}}(\boldsymbol{q}_{\mathrm{loc}},\boldsymbol{V}_{\mathrm{loc}}^{i})$, $S_{\mathrm{rel}}=f_{\mathrm{m}}(\boldsymbol{q}_{\mathrm{rel}},\boldsymbol{V}_{\mathrm{rel}}^{i})$. 最终，当前候选物体的匹配分数计算为

$$S_{\mathrm{o}}=w_{\mathrm{s}}S_{\mathrm{sub}}+w_{\mathrm{l}}S_{\mathrm{loc}}+w_{\mathrm{r}}S_{\mathrm{rel}}. \tag{8}$$

2.5 损失函数计算

本文使用三重排序损失（triplet ranking loss）来促使模型为正确的区域分配更高的分数，并减少错误区域的分数：

$$L_{\mathrm{rank}}=\sum_{i}([m-S(R^{i},E^{i})+S(R^{i},E^{j})]+[m-S(R^{i},E^{i})+S(R^{j},E^{i})]). \tag{9}$$

除此之外，由于属性在区分同类对象中起着重要作用，与 Yu 等人[1]相同，本文也增加了属性学分支来更好地理解表达式中相应的描述. 属性识别使用的标签是使用现成的语言解析器从表达式中提取的基本属性. 属性分支被定义为从 $\boldsymbol{v}^{i}$ 出发，预测区域所具有的属性特点. 在属性分支训练的过程中，本文使用了二分类交叉熵损失函数（binary-cross-entropy loss）来进行多标签分类训练. 损失函数值被定义为 R_i 第 k 个属性的预测概率 p_{ik} 与真实值 y_{ik} 之间的差距：

$$L_{\mathrm{attr}}=\sum_{i}\sum_{k}\ln(p_{ik})+(1-y_{ik})\ln(1-p_{ik}). \tag{10}$$

最终模型的总体损失函数被定义为

$$L=L_{\mathrm{attr}}+L_{\mathrm{rank}}. \tag{11}$$

2.6 擦除策略

指代短语通常从多个角度来描述视觉内容，可以为 REC 提供多种线索，如类别、属性以及与其他对象的关系等. 但是常用的注意力机制往往只捕捉最具辨别力的信息[17]. 为了克服这一局限性，Liu 等人[14]使用了一种注意力引导的擦除方法来擦除最主要的内容，利用擦除后的数据进行训练，鼓励模型发现其他补充线索以找到正确的区域.

擦除的核心是找到最主要的内容，并擦除它们. 对于指代短语理解任务，需要擦除的内容是文本中描述视觉内容的多个角度中的一个. 对于视觉特征，Liu 等人[14]根据学习到的注意权重，对重要网格的视觉特征、位置特征或邻近区域的特征进行擦除，是非常全面且有效的. 而对于指代短语，权重最高的单词

被替换为占位符单词“<UKN>”,但是由于一个视角的描述通常由几个连续的词组成,仅替换一个词并不能完全掩盖最具辨别力的“角度”.以图 3 为例,短语包含 2 个用于定位正确区域的“角度”,即“white sweater”(白色毛衣)和“polka dot skirt”(圆点裙).上面一行为 Liu 等人[14]的擦除策略和学习到的视觉注意力结果,下面一行为本文方法的结果.当只有“sweater”被替换为“<UKN>”时,这个线索仍然可以提供“white”的线索.如图 3 的右上热图所示,视觉注意力会转而优先考虑白色区域.显然,促使模型发现和利用“polka dot skirt”(圆点裙)的目的并没有被很好地实现.

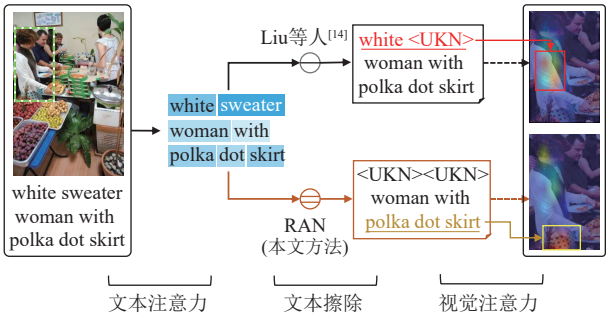


Fig. 3 Illustration of Liu et al's and our erase strategies

图 3 Liu 等人和本文的文本擦除策略示意

为了解决上述问题,本文将文本擦除的范围从一个单词扩展到多个连续的单词,这些单词可以是对视觉特征在一个角度上的完整描述.本文使用权重最高的词来寻找最具辨别力的线索的大致位置,然后通过擦除该词语及其周围的词来擦除对于该角度的完整描述.本文方法将线索的最大字数设置为 2,对于输入短语 E, x_t 代表注意力权重最高的单词,权重为 a_t.如果 x_t 为名词或形容词,则很有可能出现图 3 的情况,此时就需要在 x_t 周围寻找第 2 个待擦除的单词.本文计算要擦除的第 2 个单词的权值阈值 $\tau=(1-a_t)\cdot\sigma$,其中 σ 是控制该阈值的参数,具体值根据实验来选择.除了 x_t 之外,如果 x_{t-1} 和 x_{t+1} 的权重高于阈值 τ,本文就擦除其中权重更高的单词.如图 3 所示,当本文擦除“white sweater”时,模型只能利用“polka dot skirt”寻找正确区域,学习到的视觉注意力图会相应地对相关区域分配更高的权重.

3 实 验

本节首先介绍实验使用的数据集和一些实现细节;然后,将本文的方法与现有的指代短语理解方法进行了比较,并通过消融实验来验证模型中不同组成部分的有效性;最后,本文在可视化中展示了一些实例,直观地展示方法效果.

3.1 数据集

如表 1 和表 2 所示,本文在 3 个常用数据集上进行实验:RefCOCO[52], RefCOCO+[52], RefCOCOg[19].所有数据集所用的图像都是从 MSCOCO[53]中收集来的.RefCOCO 和 RefCOCO+数据集是通过交互式游戏收集的.RefCOCOg 是在非交互设置中注释的,指代短语比其他数据集长(平均长度为 8.43 个单词).RefCOCO 包含 142 210 个指代短语,用来描述 19 994 个图像中的 50 000 个对象;RefCOCO+在 19 992 个图像中有 49 856 个对象和 141 564 个指代短语.RefCOCOg 包含 25 799 个图像和 95 010 个指代短语,共有 49 822 个被指代的对象.RefCOCO 和 RefCOCO+中的每个图像至少包含 2 个相同类别的对象.在 RefCOCO+中禁止使用像“left”这样的绝对位置词,每张图像中有 2~4 个同类物体.

RefCOCO 和 RefCOCO+被划分为“train”“val”“testA”“testB”四个集合.“train”集合是用来训练本文的模型,“val”“testA”“testB”分别用于在不同的角度评估模型.“testA”中的图像包含多人,“testB”中的图像包含多个其他类别的对象.图像中的对象不会在不同的分割之间重叠.RefCOCOg 有 2 种类型的划分:第 1 种类型[19]将对象随机分为“train”“val”“testA”,由于测试集未发布,因此最近的工作将在验证集上进行评估,此验证集表示为“val*”.因为该划分是基于对象的,所以相同的图像可能同时出现在“train”“val”集合中.在第 2 种划分方式中[20],图像分为“train”“val”“testA”集合,本文的实验采用“val”和“test”进行数据划分.

3.2 实现细节

用于提取候选物体特征表示的 Faster R-CNN 是在 MSCOCO 的测试集中预训练的,出现在 3 个数据集的验证集和测试集中的图像没有参与预训练过程.采用的 dropout 比率为 0.1, Faster R-CNN 的参数在特征抽取过程中是固定的(这部分的实验设置与文献[1]相同,因此没有在实验部分进行更多验证).在数据规范化预处理中,与文献[1]相同,本文为每个标注物体保留最多 3 个短语,指代短语的单词数 m 被设置为每一个批次中最大的长度,为候选物体选择的上下文物体的个数被设置为 5,对于那些不够 5 个上下文物体的情况,用全 0 向量进行补齐.短语的词向量维度 d_w=300,抽取短语特征的双向 LSTM 的隐

Table 1 Comparison with the State-of-the-Art Methods for REC Based on Ground-truth Regions

表 1 与现有的基于真实区域的指代短语理解方法的比较 %

方法	特征抽取的网络	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val*	val	test
MMI[19]	VGG16		71.72	71.09		58.42	51.23	62.14		
visdif[21]	VGG16		67.57	71.19		52.44	47.51	59.25		
visdif+MMI[21]	VGG16		73.98	76.59		59.17	55.62	64.02		
NegBag[20]	VGG16	76.90	75.6	78.00						68.40
Speaker[22]	VGG16	79.56	78.95	80.22	62.26	64.6	59.62		72.63	
Attr[43]	VGG19		78.85	78.07		61.47	57.22		69.83	
VC[44]	VGG16		78.98	82.39		62.56	62.9		73.98	
A-ATT[38]	VGG16	81.27	81.17	80.01	65.56	68.76	60.63		73.18	
MattNet[1]	VGG16	80.94	79.99	82.3	63.07	65.04	61.77	73.08	73.04	72.79
CMRIN[33]	VGG16	84.02	84.51	82.59	71.46	75.38	64.74		76.16	76.25
DGA[11]	VGG16	83.73	83.56	82.51	68.99	72.72	62.98		75.76	75.79
LGRANs[31]	VGG16	82.00	81.20	84.00	66.6	67.6	65.5		75.4	74.7
MattNet[1]	ResNet101	85.65	85.26	84.57	71.01	75.13	66.17		78.1	78.12
CM-Att[14]	ResNet101	86.23	86.57	85.36	72.36	74.64	67.07		78.68	78.58
CM-Att-Erase[14]	ResNet101	87.47	88.12	86.32	73.74	77.58	68.85		80.23	80.37
CMRIN[33]	ResNet101	86.99	87.63	84.73	**75.52**	**80.93**	68.99		80.45	80.66
NMTREE[34]	ResNet101	85.65	85.63	85.08	72.84	75.74	67.62	78.03	78.57	78.21
SGMN[12]	ResNet101		86.67	85.36		78.66	69.77			81.42
Zhang 等人[45]	ResNet101	85.81	86.38	84.5	72.48	75.85	67.14		79.74	79.32
DGA[11]	ResNet101	86.34	86.64	84.79	73.56	78.31	68.15		80.21	80.26
RAN（本文）	ResNet101	**87.92**	**88.77**	**87.16**	74.67	78.92	**69.94**		**80.90**	**81.85**

注：testA 集合主要针对人的短语描述，testB 集合针对的是其他对象. 黑体数值表示最优值.

含层向量以及经过线性变换后的位置向量维度 d_q 和 d_v 都被设置为 512. 本文使用 Adam 算法来训练模型，学习率（learning rate）被初始化为 0.000 1. 本文方法建立在经典的 MattNet[1] 基础上. 在模型训练中，本文的所有实验均在显存为 13 GB 的 NVIDIA 2080Ti 显卡中训练，训练策略与 Liu 等人[14] 的相同，首先在原始数据上预训练提出的 RAN 模型（所有模型均训练 15 个 epoch），然后在擦除后的数据中对训练好的模型进行微调（训练 30 个 epoch）. 为了更好地适应比较短的指代短语，本文将擦除的单词长度设置为 2，即：对于长度大于 3 的短语，最多擦除短语中的 2 个单词.

3.3 与现有方法对比

给定指代短语 E，RAN 为图像中的每一个候选区域 $\{R_i\}_{i=1}^{N}$ 计算匹配分数，最终选择匹配分数最高的区域作为定位到的区域，评价指标为准确率. 本文在基于真实区域和基于检测区域 2 种实验设置下与现有的指代短语理解方法进行了比较，实验结果分别展示在表 1 和表 2 中. 在基于真实区域的实验中，候选区域来自 MSCOCO 的原始标注，正确的区域包含在候选区域中，因此，只有当正确的区域被选择时，才认为模型得到了正确的结果. 在基于检测区域的设置中，候选区域为利用 Faster R-CNN 检测的区域，当模型选择的区域与真实区域的面积交并比（intersection-over-union, IoU）大于 0.5 时，则认为模型给出的结果是正确的.

如表 1 和表 2 所示，在 2 种实验设置下，本文的方法都得到了最佳的性能. 整体上，以 ResNet101 为基础网络的模型效果高于以 VGGNet 为基础网络的模型，这说明更强大的特征抽取器可以帮助模型更好地理解图像内容特征，有利于分析物体属性. 与典型的模块化网络 MattNet[1] 相比，RAN 使用的基于注意的关系聚合方法能够提取更重要的内容，因此匹配分数与该精确信息相比更准确，从而预测出正确的区域. 此外，基于图结构[11,31,33] 的方法考虑了各区域之间的复杂关系和表达式之间的关系，模型关系是全面的，但是理解起来也更加复杂；而本文的方法

Table 2 Comparison with the State-of-the-Art Methods for REC Based on Automatically Detected Regions

表 2 与现有的基于自动检测区域的指代短语理解方法的比较 %

方法	特征	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val*	val	test
MMI[19]	VGG16		64.90	54.51		54.03	42.81	45.85		
NegBag[20]	VGG16		58.60	56.40				39.50		
CMN[13]	VGG16		71.03	65.77		54.32	47.76	57.47		
Speaker[22]	VGG16		72.88	63.43		60.43	48.74		59.51	
Attr[43]	VGG19		72.08	57.29		57.97	46.2		52.35	
VC[44]	VGG16		73.33	67.44		58.40	53.18		62.30	
LGRANs[31]	VGG16		76.60	66.40		64.00	53.4			62.50
MattNet[1]	ResNet101	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27	
CM-Att-Erase[14]	ResNet101	78.35	83.14	71.32	68.09	73.65	58.03		67.99	68.67
NMTREE[34]	ResNet101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
DGA[11]	ResNet101		78.42	65.53		69.07	51.99			63.28
Ref-NMS[46]	ResNet101	80.70	84.00	76.04	68.25	73.68	59.42		70.55	70.62
Sun 等人[47]	ResNet-101		74.27	68.10		71.05	58.25			70.05
RCCF（*）[35]	DLA-34		81.06	71.85		70.35	56.32			65.73
SSG（*）[48]	Darknet53		76.51	67.5		62.14	49.27	47.47	58.80	
One-Stage（*）[36]	Darknet53	72.05	74.81	67.59	55.72	60.37	48.54	48.14	59.03	58.70
ReSC（*）[49]	DarkNet53	76.59	78.22	73.25	63.23	66.64	55.53	60.96	64.87	64.87
MCN（*）[50]	DarkNet-53	80.08	82.29	74.98	67.16	72.86	57.31		66.46	66.00
LBYLNet（*）[51]	DarkNet-53	79.67	82.91	74.15	68.64	73.38	59.49		62.70	
RAN（本文）	ResNet101	78.97	83.76	72.13	68.84	74.28	58.63		68.45	69.77
RAN+Ref-NMS（本文）	ResNet101	**80.96**	**84.16**	**76.21**	**69.72**	**74.66**	**59.69**		**71.16**	**70.79**

注：testA 集合主要针对人的短语描述，testB 集合针对的是其他对象. “（*）”代表单阶段方法，直接从图像中定位被指代区域. 黑体数值表示最优值.

只考虑候选区域与其邻域之间的有用关系，消除了整个图的复杂搜索，从而更方便地聚合了有效信息. 另外 Ref-NMS[46] 是一种针对可插拔的候选物体筛选方法，可以应用在现有的基于检测的实验设置中，为了更公平地比较，本文也将其用在了 RAN 中，如表 2 所示，本文的方法取得了最佳的效果.

3.4 指代短语分割结果

本文在表 3 中显示了 RAN 在指代短语分割（referring expression segmentation, RES）[54-55] 中的结果. 与 Yu 等人[1] 一致，本文首先用训练好的 RAN 模型输出与指代短语匹配分数最高的区域，然后本文直接用预测的区域计算语义分割图. 本文使用 P@0.5 和交并比（IoU）作为评价指标. P@0.5 表示预测的指代短语分割图与真实的分割图的交并比至少为 0.5. 如表 3 所示，与现有方法相比，本文提出的 RAN 在 2 个评价指标中都取得了最好的效果.

3.5 消融实验

本节进行了一系列消融实验以说明本文方法中每个模块的有效性，相关结果见表 4，由于 RefCOCO+ 和 RefCOCOg 这 2 个数据集的实验结果趋势类似，表 4 中只展示了来自 RefCOCO 数据集的结果. 本文的基线模型是 MattNet[1]，它包含主题、位置和关系模块. 每个模块的匹配分数是根据图像和表达式中特征的连接来计算的. “RC”是 3.2 节中描述的关系构建模块，“RA”是 3.3 节中所示的关系聚合模块，“Erase”为 3.6 节中说明的擦除方法. “√”和“×”表示相应的模块是否有被使用；“RC-‡”表示在进行关系构建时，只考虑上下文-候选关系；“Erase-‡1”表示使用 Liu 等人[14] 提出的擦除方式，即每次只擦除权重最高的 1 个单词；“Erase-‡3”表示擦除单词的个数设置为 3 时的效果. 表 4 中最后一行为本文提出的 RAN 的效果，同时考虑上下文-候选关系和上下文物体之间的关系，在擦除的过程中最多擦除 2 个单词.

从表 4 可以得出 3 条结论：1）“RC-‡”效果略好于基线模型，说明注意力机制在 REC 中是有用的. 2）由于同时考虑了多个上下文区域之间的关系和上下

Table 3 Comparison with the State-of-the-Art RES Methods

表 3 与现有的指代短语分割方法的比较 %

评价指标	方法	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
P@0.5	D+RMI+DCRF[54]	42.99	42.99	44.99	20.52	21.22	20.78		
	MattNet[1]	75.16	79.55	68.87	64.11	70.12	54.82	64.48	65.60
	Chain[34]	73.36	77.55	67.30	61.60	67.15	52.24	59.64	60.29
	NMTREE[34]	74.71	79.71	68.93	65.06	70.24	56.15	63.77	64.63
	RAN（本文）	**76.76**	**81.23**	**70.17**	**65.28**	**70.75**	**56.65**	**65.34**	**66.72**
IoU	D+RMI+DCRF[54]	45.18	45.69	45.57	29.86	30.48	29.50		
	MattNet[1]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
	Chain[34]	55.29	60.99	51.36	44.74	49.83	38.50	42.55	43.99
	NMTREE[34]	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
	RAN（本文）	**58.16**	**64.39**	**53.18**	**48.10**	**53.15**	**41.63**	**47.98**	**49.32**

注：黑体数值表示最优值.

Table 4 Results of Ablation Study on RefCOCO Dataset

表 4 RefCOCO 数据集中消融实验结果

RC	RA	Erase	val	testA	testB
×	×	×	85.69	85.30	85.04
‡	√	×	86.20	86.57	85.29
√	√	×	86.72	86.99	86.01
√	√	‡1	87.51	88.29	86.50
√	√	‡3	87.47	88.43	86.69
√	√	√	87.92	88.77	87.16

注：“√”和“×”表示相应的模块是否有被使用，“‡”表示模块用不同的方式实现.

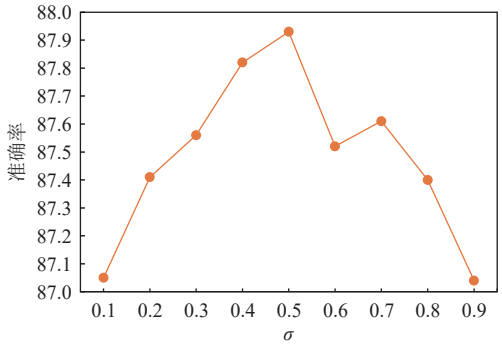


Fig. 4 Performance on the validation split of RefCOCO when σ is set as different values

图 4 σ 设置为不同值时 RefCOCO 的验证集效果

文候选关系，同时使用关系构建模块和关系聚合模块进一步提高了性能. 3)如表 4 最后 3 行所示，与其他 2 种擦除方式相比，本文的擦除策略可以更好地测试模型、发现更多的信息，从而提高模型效果，相对于 Liu 等人[14]的擦除方式，本文的方法更能完整地擦除短语在一个角度的描述；当擦除的单词数设置为 3 时，需要短语的单词数大于 4，而满足条件的短语数较少，这影响了擦除策略的效果，正因如此，本文的实验部分将擦除单词的最大个数限制为 2.

3.6 参数敏感性分析

本节测试了擦除第 2 个单词采用的阈值中参数 σ 对于模型实验效果的影响. 图 4 展示了当 σ 取 0.1~0.9 时 RAN 在 RefCOCO 上的准确率的变化情况. 如图 4 所示，当 σ=0.5，也就是阈值 $\tau = (1 - \alpha_t)\times 0.5$ 时，模型效果最好.因此本文其他部分实验中 σ 的值都被设置为 0.5.

由于 τ 控制的是第 2 个单词擦除与否，在一定程度上代表了文本注意力机制是否为第 2 个单词分配了更高的权重. 当注意力权重同时为连续的 2 个单词都分配了较高的权重，说明这 2 个单词大概率具有连贯的语义，例如图 3 中的“white sweater”. σ 值较小代表放宽连续 2 个单词语义连贯性的要求，同时擦除 2 个单词的数据比例更高，对模型的微调提出更高的要求，一方面迫使模型发现更多有用信息，另一方面也可能会影响模型对于整个短语的理解从而限制模型的效果. 而 σ 值较大时效果恰好相反，σ=0.5 恰好是 2 方面的折中，因而具有最佳效果.

3.7 定性结果

本文在图 5 中提供了定性结果，它显示了学习到的候选区域的视觉注意力结果以及对预测结果贡献最大的上下文区域. 第 1 行是 MattNet[1]的结果，第 2 行是本文 RAN 结果. 对于图 5(a)的示例，基线模型和 RAN 都能根据短语中描述的物体类别和属性信息找到正确的区域. 但是对于图 5(b)中的示例，尽管基

(a) 输入: kid in black　(b) 输入: man in black on the bike　(c) 输入: red guy on the left next to 3 others　(d) 输入: man standing alone

注：单实线框和虚线框分别代表正确预测的结果和错误预测的结果，双实线框代表对结果预测贡献最大的上下文区域.

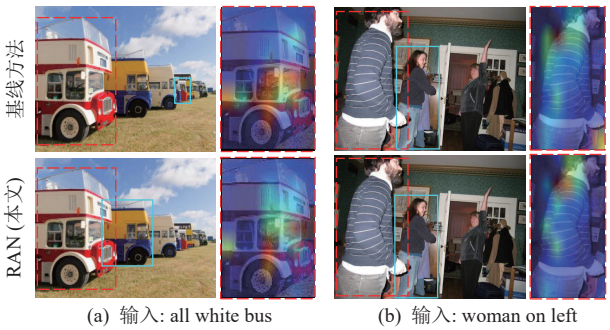
Fig. 5　Visualization of MattNet[1] and RAN on RefCOCO dataset

图 5　MattNet[1] 和 RAN 在 RefCOCO 数据集的预测结果可视化

线模型预测出了正确的区域，但却选择了错误的上下文，得到的正确预测结果来自于区域属性和类别. 相比之下，由于 RAN 充分考虑了候选区域和上下文区域之间的关系，因此在准确的上下文的基础上得到了正确的结果. 图 5(c)中的示例与图 1 中的示例相同，要求充分理解多个对象之间的关系，找到正确的人需要理解“3 others”的含义，并在上下文区域中找到他们. RAN 能够同时对上下文-候选关系和多个上下文区域之间的关系进行建模，因此能够从综合关系中预测正确的区域. 此外，图 5(d)中的指代短语涉及由“alone”一词隐含地描述与多个对象的关系，基于正确的上下文信息，RAN 也得到了正确的结果.

3.8　错误样例分析

图 6 展示了一些失败案例. 由于输入的文本短语和视觉内容之间存在语义鸿沟，很难在图像中准确反映语言线索. 如图 6 所示，白色部分在学习的视觉注意中被突出显示，但是表达中的完整线索是“all white”(全白色). 此外，如图 6(b)所示，该短语提到细粒度的类别“woman”(女性)，在特征抽取器的预训练过程中并没有使用“woman”，基线模型和本文的方法都理解了预测区域是“man”(人)，而不是更加细粒度的类别“man”(男人)和“woman”(女人). 因此，模型很难根据表达式中的类别来区分具有相同视觉类别的 2 个区域，从而导致错误的预测. 这个问题在其他跨模态任务中也很普遍，我们将在今后的工作中努力解决这个问题.



(a) 输入: all white bus　(b) 输入: woman on left

Fig. 6　Failure cases

图 6　失败样例

4　总　　结

本文提出了一个关系聚合网络(RAN)来完成指代短语理解的任务.RAN 模型根据完备的关系建立了表达区域和其他图像区域之间的映射关系. 具体地说，RAN 构建了上下文候选关系和多个上下文区域之间的关系，利用图像与指代短语之间的跨模态对应关系来聚合所构建的关系. 此外，本文还设计了一种新的擦除策略，促使模型根据更多的线索来预测正确的区域. 在 3 个数据集上的实验结果表明了该方法的优越性.

作者贡献声明：郭文雅提出了算法的思路和实验方案；张莹完成了论文初稿撰写；刘胜哲完成论文相关实验；杨巨峰对论文内容进行润色；袁晓洁确定了论文最终内容.

参　考　文　献

[1] Yu Licheng, Lin Zhe, Shen Xiaohui, et al. MattNet: Modular attention network for referring expression comprehension[C] //Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1307−1315

[2] Meng Xiangshen, Jiang Aiwen, Liu Changhong, et al. Visual question

answering based on spatial DCTHash dynamic parameter network[J]. SCIENTIA SINICA Informations, 2017, 47(8): 60−74 (in Chinese)
(孟祥申, 江爱文, 刘长红, 等. 基于Spatial-DCTHash 动态参数网络的视觉问答算法[J]. 中国科学: 信息科学, 2017, 47(8): 60−74)

[3] Li Guohao, Wang Xin, Zhu Wenwu. Boosting visual question answering with context-aware knowledge aggregation[C] //Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 1227−1235

[4] Zhou Yiyi, Ji Rongrong, Sun Xiaoshuai, et al. *K*-armed bandit based multi-modal network architecture search for visual question answering[C] //Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 1245−1254

[5] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C] //Proc of the 36th Int Conf on Machine Learning. Piscataway, NJ: IEEE, 2015: 2048−2057

[6] Zhang Beichen, Li Liang, Su Li, et al. Structural semantic adversarial active learning for image captioning[C] //Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 1112−1121

[7] Wang Yong, Zhang Wenkai, Liu Qing, et al. Improving intra- and inter-modality visual relation for image captioning[C] //Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 4190−4198

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409.1556, 2014

[9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C] //Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[10] Liu Yongfei , Wan Bo, Zhu Xiaodan, et al. Learning cross-modal context graph for visual grounding[C] //Proc of the 34th Association for the Advancement of Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11645−11652

[11] Yang Sibei, Li Guanbin, Yu Yizhou. Dynamic graph attention for referring expression comprehension[C] //Proc of the 17th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4643−4652

[12] Yang Sibei, Li Guanbin, Yu Yizhou. Graph-structured referring expression reasoning in the wild[C] //Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9949−9958

[13] Hu Ronghang, Rohrbach M, Andreas J, et al. Modeling relationships in referential expressions with compositional modular networks[C] // Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4418−4427

[14] Liu Xihui, Wang Zihao, Shao Jing, et al. Improving referring expression grounding with cross-modal attention-guided erasing[C] // Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 1950−1959

[15] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint, arXiv: 1710.10903, 2017

[16] Xu Jinghang, Zuo Wanli, Liang Shining, et al. Causal relation extraction based on graph attention networks[J]. Journal of Computer Research and Development, 2020, 57(1): 159-174 (in Chinese)
(许晶航, 左万利, 梁世宁, 等. 基于图注意力网络的因果关系抽取[J]. 计算机研究与发展, 2020, 57(1): 159-174)

[17] Zhang Xiaolin, Wei Yunchao, Feng Jiashi, et al. Adversarial complementary learning for weakly supervised object localization[C] // Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1325−1334

[18] Luo Ruotian, Shakhnarovich G. Comprehension-guided referring expressions[C] //Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3125−3134

[19] Mao Junhua, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions[C] //Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 11−20

[20] Nagaraja V K, Morariu V I, Davis L S. Modeling context between objects for referring expression understanding[C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 792−807

[21] Yu Licheng, Poirson P, Yang Shan, et al. Modeling context in referring expressions[C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 69−85

[22] Yu Licheng, Tan Hao, Bansal M, et al. A joint speaker-listener-reinforcer model for referring expressions[C] //Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3521−3529

[23] Fukui A, Park D H, Yang D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[C] //Proc of the 21st Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 457−468

[24] Rohrbach A, Rohrbach M, Hu Ronghang, et al. Grounding of textual phrases in images by reconstruction[C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 817−834

[25] Andreas J, Rohrbach M, Darrell T, et al. Neural module networks[C] // Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 39−48

[26] Hu Ronghang, Andreas J, Rohrbach M, et al. Learning to reason: End-to-end module networks for visual question answering[C] //Proc of the 16th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 804−813

[27] Xian Guangjing, Huang Yongzhong. A survey of visual question answering technology based on neural network[J]. Network Security Technology & Application, 2018, 1: 42−47 (in Chinese)
(鲜光靖, 黄永忠. 基于神经网络的视觉问答技术研究综述[J]. 网络安全技术与应用, 2018, 1: 42−47)

[28] Johnson J, Hariharan B, Van D M L, et al. Inferring and executing programs for visual reasoning[C] //Proc of the 16th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 3008−3017

[29] Du Pengfei, Li Xiaoyong, Gao Yali, et al. Survey on multimodal visual language representation learning[J]. Journal of Software, 2021, 32(2): 327−348 (in Chinese)
(杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述[J]. 软件学报, 2021, 32(2): 327−348)

[30] Andreas J, Klein D, LevinE S. Modular multitask reinforcement

learning with policy sketches[C] //Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2017: 166−175

[31] Wang Peng, Wu Qi, Cao Jiewei, et al. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks[C] //Proc of the 37th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 1960−1968

[32] Bajaj M, Wang Lanjun, Sigal L. G^{3}raphground: Graph-based language grounding[C] //Proc of the 17th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4280−4289

[33] Yang Sibei, Li Guanbin, Yu Yizhou. Cross-modal relationship inference for grounding referring expressions[C] //Proc of the 37th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4145−4154

[34] Liu Daqing, Zhang Hanwang, Zha Zhengjun, et al. Learning to assemble neural module tree networks for visual grounding[C] //Proc of the 17th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4672−4681

[35] Liao Yue, Liu Si, Li Guanbin, et al. A real-time cross-modality correlation filtering method for referring expression comprehension[C] //Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10877−10886

[36] Yang Zhengyuan, Gong Boqing, Wang Liwei, et al. A fast and accurate one-stage approach to visual grounding[C] //Proc of the 17th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4682−4692

[37] Yu Zhou, Yu Jun, Xiang Chenchao, et al. Rethinking diversified and discriminative proposal generation for visual grounding[C] //Proc of the 27th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 1114−1120

[38] Deng Chaorui, Wu Qi, Wu Qingyao, et al. Visual grounding via accumulated attention[C] //Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7746−7755

[39] Hu Zhiwei, Feng Guang, Sun Jiayu, et al. Bi-directional relationship inferring network for referring image segmentation[C] //Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 4423−4432

[40] Zhuang Bohan, Wu Qi, Shen Chunhua, et al. Parallel attention: A unified framework for visual object discovery through dialogs and queries[C] //Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4252−4261

[41] Ren Shaoqing, He Kaiming, Girshick R B, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2017, 39(6): 1137−1149

[42] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C] //Proc of the 19th Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1532−1543

[43] Liu Jingyu, Wang Liang, Yang Ming-Hsuan. Referring expression generation and comprehension via attributes[C] //Proc of the 16th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 4866−4874

[44] Zhang Hanwang, Niu Yuelei, Chang S. Grounding referring expressions in images by variational context[C] //Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4158−4166

[45] Zhang Chao, Li Weiming, Ouyang Wanli, et al. Referring expression comprehension with semantic visual relationship and word mapping[C] //Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1258−1266

[46] Chen Long, Ma Wenbo, Xiao Jun, et al. Ref-NMS: Breaking proposal bottlenecks in two-stage referring expression grounding[C] //Proc of the 35th Association for the Advancement of Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 1036−1044

[47] Sun Mingjie, Xiao Jimin, Lim E G. Iterative shrinking for referring expression grounding using deep reinforcement learning[C] //Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14060−14069

[48] Chen Xinpeng, Ma Lin, Chen Jingyuan, et al. Real-time referring expression comprehension by single-stage grounding network[J]. arXiv preprint, arXiv: 1812.03426, 2018

[49] Yang Zhengyuan, Chen Tianlang, Wang Liwei, et al. Improving one-stage visual grounding by recursive sub-query construction[C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 387−404

[50] Luo Gen, Zhou Yiyi, Sun Xiaoshuai, et al. Multi-task collaborative network for joint referring expression comprehension and segmentation[C] //Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10031−10040

[51] Huang Binbin, Lian Dongze, Luo Weixin, et al. Look before you leap: Learning landmark features for one-stage visual grounding[C] //Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 16888−16897

[52] Kazemzadeh S, Ordonez V, Matten M, et al. Referitgame: Referring to objects in photographs of natural scenes[C] //Proc of the 19th Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 787−798

[53] Lin T, Maire M, Belongie S J, et al. Microsoft COCO: Common objects in context[C] //Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 740−755

[54] Liu Chenxi, Lin Zhe, Shen Xiaohui, et al. Recurrent multimodal interaction for referring image segmentation[C] //Proc of the 16th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1280−1289

[55] Hu Ronghang, Rohrbach M, Darrell T. Segmentation from natural language expressions[C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 108−124

Guo Wenya, born in 1994. PhD. Her main research interests include multimodal data processing and sentiment analysis.

郭文雅, 1994年生. 博士. 主要研究方向为多模态数据处理和情感分析.

Zhang Ying, born in 1986. PhD, professor, PhD supervisor. Her main research interests include natural language processing, sentiment analysis, and multimodal data analysis.

张　莹, 1986 年生. 博士, 教授, 博士生导师. 主要研究方向为自然语言处理、情感分析、多模态数据分析.

Liu Shengzhe, born in 1998. Master. His main research interest includes weakly supervised visual grounding.

刘胜哲, 1998 年生. 硕士. 主要研究方向为弱监督视觉文本定位.

Yang Jufeng, born in 1980. PhD, professor, PhD supervisor. His main research interests include visual sentiment analysis, fine-grained classification, medical image recognition, and image retrieval.

杨巨峰, 1980 年生. 博士, 教授, 博士生导师. 主要研究方向为视觉情感计算、细粒度分类、医疗图像识别和图像检索.

Yuan Xiaojie, born in 1963. PhD, professor, PhD supervisor. Her main research interests include big data analysis, data mining, and database technology.

袁晓洁, 1963 年生. 博士, 教授, 博士生导师. 主要研究方向为大数据分析、数据挖掘和数据库技术.