控制与决策
Control and Decision

一种基于多层语义特征的图像理解方法

莫宏伟， 田朋

引用本文:
莫宏伟, 田朋. 一种基于多层语义特征的图像理解方法[J]. 控制与决策, 2021, 36(12): 2881−2890.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0927

您可能感兴趣的其他文章

Articles you may be interested in

基于多层级特征的机械臂单阶段抓取位姿检测

Single−stage grasp pose detection of manipulator based on multi−level features

控制与决策. 2021, 36(8): 1815−1824 https://doi.org/10.13195/j.kzyjc.2019.1840

基于多尺度特征表示的行人再识别

Multi−scale feature representation for person re−identification

控制与决策. 2021, 36(12): 3015−3022 https://doi.org/10.13195/j.kzyjc.2020.0952

改进YOLOv2的端到端自然场景中文字符检测

End−to−end Chinese character detection in natural scene based on improved YOLOv2

控制与决策. 2021, 36(10): 2483−2489 https://doi.org/10.13195/j.kzyjc.2020.0270

结合注意力机制的循环神经网络复述识别模型

Recurrent neural networks based paraphrase identification model combined with attention mechanism

控制与决策. 2021, 36(1): 152−158 https://doi.org/10.13195/j.kzyjc.2019.0638

基于联合知识表示学习的多模态实体对齐

Multi−modal entity alignment based on joint knowledge representation learning

控制与决策. 2020, 35(12): 2855−2864 https://doi.org/10.13195/j.kzyjc.2019.0331

一种基于多层语义特征的图像理解方法

莫宏伟, 田 朋†

(哈尔滨工程大学 智能科学与工程学院，哈尔滨 150001)

摘 要: 视觉场景理解包括检测和识别物体、推理被检测物体之间的视觉关系以及使用语句描述图像区域.为了实现对场景图像更全面、更准确的理解,将物体检测、视觉关系检测和图像描述视为场景理解中3种不同语义层次的视觉任务,提出一种基于多层语义特征的图像理解模型,并将这3种不同语义层进行相互连接以共同解决场景理解任务.该模型通过一个信息传递图将物体、关系短语和图像描述的语义特征同时进行迭代和更新,更新后的语义特征被用于分类物体和视觉关系、生成场景图和描述,并引入融合注意力机制以提升描述的准确性. 在视觉基因组和COCO数据集上的实验结果表明,所提出的方法在场景图生成和图像描述任务上拥有比现有方法更好的性能.

关键词: 图像理解；语义层；语义特征；视觉关系；场景图；图像描述；注意力机制

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0927 **开放科学(资源服务)标识码(OSID):**

引用格式: 莫宏伟,田朋. 一种基于多层语义特征的图像理解方法[J]. 控制与决策, 2021, 36(12): 2881-2890.

An image understanding method based on multi-level semantic features

MO Hong-wei, TIAN Peng†

(College of Intelligent Systems Science and Engineering，Harbin Engineering University，Harbin 150001，China)

Abstract: Visual scene understanding includes detecting and recognizing objects, reasoning the visual relationships of the detected objects, and describing image regions with sentences. In order to achieve the more comprehensive and accurate understanding of scene image, we view object detection, visual relationship detection and image captioning as three visual tasks at different semantic levels in scene understanding, so as to propose an image understanding model based on multi-level semantic features to leverage the mutual connections across the three different semantic layers to solve the scene understanding tasks jointly. The model iterates and updates the semantic features of objects, relationship phrases and image captioning simultaneously through a message pass graph. The updated semantic features are used to classify objects and visual relationships, generate scene graphs and captions, and introduce a fusion attention mechanism to improve the accuracy of captions. The experimental results on the visual genome and COCO datasets show that the proposed method outperforms the existing methods on the scene graph generation and image captioning tasks.

Keywords: image understanding；semantic level；semantic feature；visual relationship；scene graph；image captioning；attention mechanism

0 引 言

近些年,场景理解已经成为计算机视觉和人工智能领域的一个研究热点,机器人通过检测和识别场景中的物体及其之间的关系等信息来理解其所处的环境[1]. 目前,物体检测和识别方法已经相对成熟,但是,理解视觉场景并不仅仅是检测和识别出单个物体,还要能够对物体之间的关系进行检测,并使用自然语言描述出场景图像所要表达的意义[2]. 相同种类的物体由于其交互关系的不同,会使得场景图像呈现出不同的内容含义. 如图1所示,虽然两幅图像都包含相同的物体:人和摩托车,但是图像表示的内容却是不同的. 理解视觉场景图像的关键不仅是将物体及其之间的交互关系进行表示,而且还要使用高级别的视觉认知任务以实现对场景图像更加全面的理解[3-4]. 现阶段只是使用物体检测识别的方法不足以实现对场景图像的全面理解,且难以感知出不同图像所表示内容的差别. 因此,设计出一种能够同时解决场景理解的多种视觉任务、准确且较为全面地表示场景图像

收稿日期: 2020-07-09；**修回日期:** 2020-09-22.
基金项目: 国家重点研发计划项目(2018AAA0102702).
责任编委: 张国山.
†通讯作者. E-mail: tianpeng0609@163.com.

所示内容的方法已成为当前的研究热点.

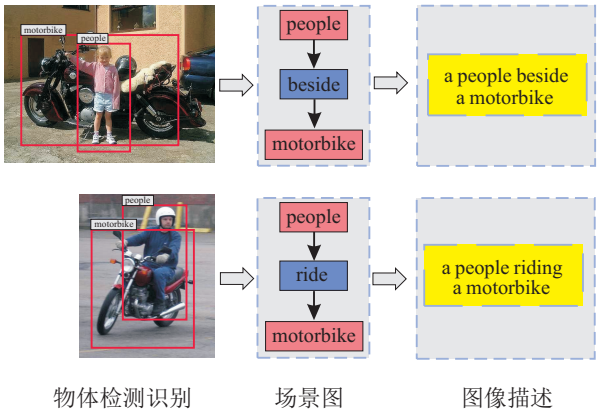


图 1 对场景图像的不同语义层进行注释

场景理解中不同的视觉任务对应着不同的语义层: 物体检测是检测图像中物体的类别和属性等信息, 侧重于单个物体, 如图 1 中的人和车; 视觉关系检测是预测物体之间的交互关系, 可将关系视为一个“主语-谓词-宾语”短语, 例如“人-骑-车”; 图像描述使用自然语言对图像所表示的主要内容进行文字描述, 其中包含多个物体及其属性和交互关系等信息.

现有的场景图像理解方法一般是从图像的单个语义层出发, 分别将物体检测识别、视觉关系检测和图像描述中的部分视觉任务作为实现场景理解的研究目标. 文献 [5] 在图像中识别物体的位置及其之间的语义关系. 文献 [6-7] 关注于目标物体之间关系的局部预测. 这些方法侧重的是解决图像理解的单个视觉任务, 并没有使用不同语义层之间的上下文特征信息以提升模型理解场景图像的性能. 文献 [8] 提出的 MSDN 模型和文献 [9] 提出的 C2SGNet 模型利用上下文信息提升了场景图生成的准确率, 但是这些方法只是提升了场景图像理解中部分视觉任务的准确率.

为了同时解决场景理解的多种视觉任务, 提升场景图像理解的准确性和全面性, 本文提出一种基于多层语义特征的图像理解方法. 通过构建一个信息传递图, 在场景理解视觉任务的不同语义层之间传递和更新特征信息, 以实现同时对物体检测、视觉关系检测和图像描述进行求解; 在图像描述的过程中引入融合注意力机制以提升描述的准确率; 模型有效地利用不同语义层的特征信息以提升图像理解的准确率. 在 Visual Genome[10] 和 COCO[11] 数据集上的实验结果表明, 本文提出的图像理解方法能够利用不同语义层之间的上下文特征提升物体检测、场景图生成和图像描述的准确率, 对于提升场景图像理解的准确性和全面性具有一定的作用.

1 相关研究

1.1 物体检测

物体检测是计算机视觉研究的热点, 也是实现场景理解的基础任务. Girshick 等[12] 首先将卷积神经网络 (convolutional neural network, CNN) 引入目标检测; 随后, 人们相继提出了多种基于 CNN 的目标检测算法以提高物体检测的准确度和效率, 如 Faster R-CNN[13]、YOLOv4[14] 和 SSD[15]. 但是, YOLO 模型没有使用区域提议机制, 导致目标位置回归精度不足; SSD 模型在区域提议和区域识别之间共享更多的卷积层, 使得网络结构变得复杂. Faster R-CNN 通过候选框提议网络 (region proposal network, RPN) 生成目标候选区域, 具有很好的检测效率. 因此, 本文使用 Faster R-CNN 作为所提出模型的基础.

1.2 视觉关系检测

视觉关系检测需要检测出图像中物体对之间的交互关系. Lu 等[16] 提出了一种利用先验语言检测视觉关系中的短语及其成分, 该方法受限于谓词或关系的数量, 准确性较差. Li 等[17] 使用主语、宾语和谓语之间的消息传递结构来对视觉关系进行建模. Xu 等[18] 通过在循环神经网络 (recurrent neural network, RNN) 中进行信息传递以迭代细化关系和物体检测. 然而, 这些方法只是单独地对视觉关系进行检测和推理, 并没有建立视觉关系与其他语义任务之间的联系. 本文将根据空间关系建立不同语义层之间的联系.

1.3 图像描述

图像描述作为场景理解的一种表示方式, 使用自然语言正确地表述图像的语义内容[19]. 图像描述方法主要包括基于模板的方法[20-21]、基于检索的方法[22-23] 和现阶段成为主流的基于编码解码结构的方法[24-26]. 在此基础上, 人们相继提出了多种基于注意力机制的图像描述模型[27-29]. 为了进一步提升图像描述的准确性, 本文提出将空间注意力机制与高层语义注意力机制进行融合的描述方法以更好地学习图像描述的特征, 同时图像描述也可以为场景理解的其他视觉任务提供有用的信息.

1.4 场景图生成

Johnson 等[30] 首先提出场景图并将其应用于图像检索任务, 场景图中的节点表示物体, 有向边表示关系. 最近, 人们提出了多种方法来阐述特征信息在物体及其之间关系的传递[17,31]. 本文关注图像中不同语义层之间的信息传递, 通过信息传递机制来细化物体和关系的表示. 现有的场景图生成方法是基于

图像的真实标注或者从其他模型中提取图形,如文献[32-33],而本文方法是通过物体检测和视觉关系检测直接从输入图像中生成场景图.

2　基于多层语义特征的图像理解模型

场景图像中包含的不同层次语义信息的相互关联和补充可以有效地提升不同语义层对应的视觉任务的准确性. 本文提出一种基于多层语义特征的图像理解模型以对场景图像中不同语义层的特征信息进行联合细化,同时解决场景理解中的物体检测、视觉关系检测、场景图生成和图像描述任务,以提升场景图像理解的准确性和全面性. 模型总体结构如图2所示.

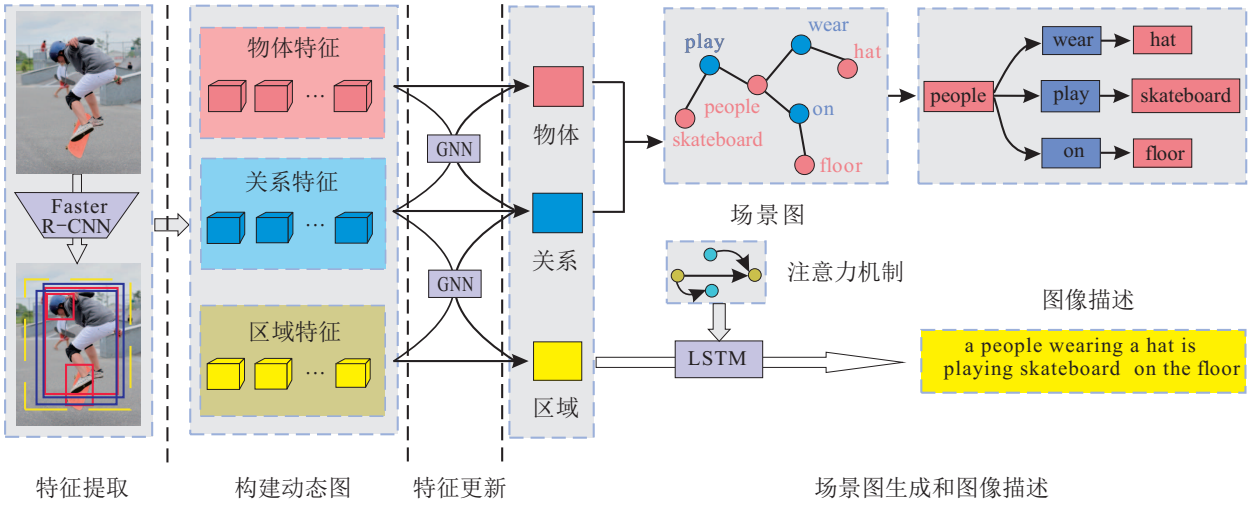


图 2　基于多层语义特征的图像理解模型的整体框图

模型的整个过程如下:

1) 区域提议和特征提取. 使用Faster R-CNN[13]检测出一组物体区域提议,将物体区域提议进行分组以生成关系区域提议,再将物体和关系区域提议连接组合形成描述区域提议,然后使用ROI (region of interest)池化分别获取各个区域的特征信息.

2) 构建信息传递图. 根据不同语义层的空间关系将不同语义层的特征进行连接.

3) 特征更新. 将信息传递图中的信息送入图神经网络(graph neural network, GNN)以实现对不同视觉任务的特征同时进行迭代更新. 更新之后的特征信息被用于物体检测识别、关系检测识别和描述图像区域的内容信息.

4) 注意力机制检测. 对图像中目标物体的空间位置和名称属性信息进行检测,使用融合空间和高层语义的注意力机制改变图像中目标物体的空间位置特征向量和名称属性向量的权重.

5) 场景图生成和图像描述. 更新后的语义特征用于对目标物体和关系谓词进行分类,随后结合注意力机制检测到的特征信息送入到长短时记忆单元(long short-term memory, LSTM)模型中生成描述语句,使用检测的目标物体和关系(谓词)生成场景图像所对应的场景图.

2.1　区域提议

对于给定的输入图像I,首先使用Faster R-CNN检测提取出一组候选区域集合$B=\{b_1,b_2,\ldots,b_n\}$,对于每一个物体区域,模型不仅提取出边界框b_i表示物体的位置,而且将使用ROI池化提取的物体空间位置特征向量x_i和名称属性向量a_i送入后续网络进行推理.

使用非极大值抑制(non-maximum suppression, NMS)对物体提议的ROI数量进行控制以减少计算量,将被检测到的N个物体提议进行连接以表示关系提议区域,将多个物体的区域提议和与其对应的关系提议进行组合以生成描述区域提议.

2.2　构建信息传递图

场景图是图像语义内容的结构化表示,场景图G可用一个定义的三元组$G=\{B,O,R\}$表示[33]:

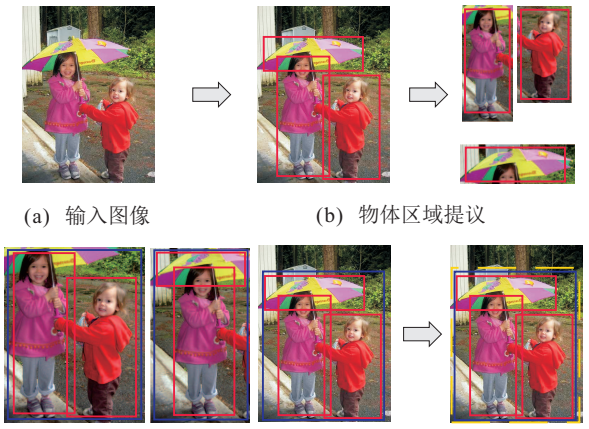
1) $B=\{b_1,b_2,\ldots,b_n\}$是候选边界框集合;

2) $O=\{o_1,o_2,\ldots,o_n\}$是物体集合,元素$o_i\in N$表示区域b_i所对应的类标签;

3) $R=\{r_{1\to2},r_{1\to3},\ldots,r_{n\to n-1}\}$是物体之间的关系集合,$r_{i\to j}$表示主体$(b_i,o_i)\in B\times O$和客体$(b_j,o_j)\in B\times O$以及关系标签$x_{i\to j}\in R$的元组.

由于不同的图像所对应场景图的拓扑结构也各不相同,基于ROI之间的空间关系建立连接图. 每一

个关系提议可看成一个“主语-谓词-宾语”三元组,并与两个物体提议进行连接,将所构造的每个描述提议与多个物体和关系提议进行连接. 图中的节点表示物体、关系和描述区域,边表示区域之间的连接. 通过上述过程,可以构造一个信息传递图以对物体区域提议、关系短语区域提议和描述区域提议之间的联系进行建模. 信息传递图如图3所示.



(a) 输入图像 (b) 物体区域提议

(c) 关系短语区域提议 (d) 描述区域提议

图 3 信息传递图

2.3 特征信息更新

场景图模型可以表示为$G=\{B,V,E\}$. 其中:B表示区域集合,V表示节点集合,E表示边集合. v^o, $v^s,v^r,v^c\in V$分别表示主语和宾语物体节点、关系节点和描述节点;$e^{s\to r},e^{r\to o}\in E$分别表示主语和宾语物体节点与关系节点之间的有向边.

本文使用GNN来实现节点的信息传递和更新,因为GNN能够记忆节点自身的特征信息,又可以将接收到的信息进行融合表示,并能够迭代更新节点表示. 在每个时间步骤,每一个节点都有一个隐藏状态,在初始时刻使用输入特征将隐藏状态初始化,将节点以前的状态和来自相邻节点的信息进行聚合以更新每个节点的表示. 由于物体对区域的信息主要用于物体之间关系的识别,将从这些物体区域提取的特征作为关系节点的输入;描述区域主要包括物体和物体对之间的关系信息,将从物体区域和关系区域提取的特征作为描述节点的输入.

对于信息传递图中的每一个物体节点,首先使用一个预先训练的检测器检测出物体区域;然后从该检测区域中提取物体特征向量$f^o\in R^d$以初始化该物体节点,否则,使用一个d维度的零向量对其进行初始化. 将从物体对区域提取的物体特征向量与位置信息一起编码成物体对的几何特征;然后,将这些特征向量送入一个全连接层以产生特征向量$f^r\in R^d$,该特征向量被作为物体对区域所对应的关系节点的输入特征;最后,将特征向量f^o和f^r作为所对应的描述节点的输入特征.

在实践中,假设$f_i^s,f_i^o\in R^D$分别是对应于v^o和v^s物体区域的特征向量,$f_i^r,f_i^c\in R^D$分别表示关系区域和描述区域的特征向量,则物体、关系和描述区域的特征信息传递更新过程如下所示.

1) 物体.

依据信息传递图,在每个时间步骤t中,每个物体节点聚合来自相邻节点(物体节点和关系节点)信息的过程可表示如下:

$$m_{is}^t=\Big[\sum_{j\in\mathcal{N}_i^s}(f_j^o,h_j^{t-1}),\sum_{j\in\mathcal{N}_i^s}(f_j^r,h_j^{t-1})\Big],\tag{1}$$

$$m_{io}^t=\Big[\sum_{j\in\mathcal{N}_i^o}(f_j^s,h_j^{t-1}),\sum_{j\in\mathcal{N}_i^o}(f_j^r,h_j^{t-1})\Big].\tag{2}$$

其中:m_{is}^t、m_{io}^t分别表示主语物体节点和宾语物体节点接收到的信息,f_j^s、f_j^o分别表示主语物体节点信息和宾语物体节点信息,f_j^r表示关系节点信息,$\mathcal{N}_i^s$、$\mathcal{N}_i^o$分别表示主语物体节点和宾语物体节点的相邻节点集合. 模型将来自其他节点的信息$m_i^t=m_{is}^t\bigcup m_{io}^t$和该节点以前时间步骤的信息输入到门控循环单元[34]以更新该节点的隐藏状态,有

$$\begin{cases}z_i^t=\sigma(W_o^z m_i^t+U_o^z h_i^{t-1}),\\ r_i^t=\sigma(W_o^r m_i^t+U_o^r h_i^{t-1}),\\ \tilde{h}_i^t=\tanh(W_o m_i^t+U_o(r_i^t\odot h_i^{t-1})),\\ h_i^t=(1-z_i^t)\odot h_i^{t-1}+z_i^t\odot\tilde{h}_i^t.\end{cases}\tag{3}$$

其中:σ表示sigmoid激活函数,$\odot$表示向量的点乘. 通过这种方式使得图中的节点能够聚合来自相邻节点的信息,同时将自身信息传播到其他相邻的节点. 通过图将节点信息进行传播并进行T次迭代后,能够得到每个物体节点所对应区域的最终隐藏状态$\{h_1^{\mathrm{T}},h_2^{\mathrm{T}},\dots,h_N^{\mathrm{T}}\}$,本文使用一个全连接网络层$\phi_o$来计算每个节点的输出特征$f_i^o$,即

$$f_i^o=\phi_o(h_i^o,h_i^{\mathrm{T}}),\tag{4}$$

将节点的初始隐藏状态和最终隐藏状态作为网络的输入.

对于每一个物体区域,将节点所有相关的输出特征聚合后送入一个物体分类全连接层δ_o中以预测物体的类别信息,预测的物体类别信息将会被用于关系的预测,有

$$l_i^o=\arg\max[\delta_o(f_1^o,f_2^o,\dots,f_N^o)].\tag{5}$$

2) 关系.

物体的分类特征有助于进行物体之间的关系检

测,在初始时刻,本文使用物体区域的特征向量来初始化物体节点,使用主语和宾语物体区域所组成的联合区域的特征向量以及物体的空间信息来初始化关系节点,有

$$h_o^0 = \varphi_o(f_i^s, f_j^o), \tag{6}$$

$$h_r^0 = \varphi_r(f_{ij}^r, s_i^s, s_j^o). \tag{7}$$

其中:φ_o和φ_r表示在全连接层上进行的转换;f_i^s、f_j^o为主语和宾语物体所对应区域的特征向量;f_{ij}^r表示主语和宾语物体区域所组成的联合区域的特征向量;s_i^s、s_j^o表示主语和宾语物体的空间信息;h_o^0表示物体节点的初始隐藏状态;h_r^0表示关系节点的初始隐藏状态,关系节点信息传递和更新方式与物体节点相同.

在每个时间步骤t中,关系节点聚合来自主语和宾语物体节点信息的过程可表示为

$$m_{ir}^t = \Big[\sum_{j\in\mathcal{N}_i^o}(f_j^s, h_j^{t-1}), \sum_{j\in\mathcal{N}_i^s}(f_j^o, h_j^{t-1})\Big]. \tag{8}$$

使用式(3)中的循环机制将该过程迭代T_r次后,获得每个节点的最终隐藏状态$\{h_1^{T_r}, h_2^{T_r}, \dots, h_M^{T_r}\}$,将初始隐藏状态和最终隐藏状态作为一个全连接网络层ϕ_r的输入来计算节点的输出特征,有

$$f_i^r = \phi_r(h_r^0, h_i^{T_r}). \tag{9}$$

对于每一个关系区域,将节点所有相关的输出特征聚合后送入到一个关系分类全连接层δ_r中用于预测关系的类别,有

$$l_{i\to j}^r = \arg\max[\delta_r(f_1^r, f_2^r, \dots, f_i^r, f_j^r, \dots, f_M^r)]. \tag{10}$$

3) 描述.

信息传递图中每个描述区域节点会连接多个物体和关系节点,其特征更新过程与上述物体和关系节点类似. 在初始时刻,使用物体区域和关系区域所组成的描述区域的特征向量来初始化对应的描述节点,有

$$h_c^0 = \varphi_c\Big[\sum_{i=1,j\in\mathcal{N}_i^s}^{N}(f_i^s, f_j^o), \sum_{i=1}^{M} f_i^r\Big]. \tag{11}$$

其中:φ_c表示在全连接层上进行的转换,h_c^0表示描述节点的初始隐藏状态. 对于描述节点信息的传递更新,本文使用与物体节点相同的方式. 在每个时间步骤t中,描述节点聚合来自于物体节点和关系节点信息的过程表示为

$$m_{ic}^t = \Big[\sum_{i=1,j\in\mathcal{N}_i^s}^{N}(f_i^s, f_j^o, h_i^{t-1}), \sum_{i=1}^{M}(f_i^r, h_i^{t-1})\Big]. \tag{12}$$

使用式(3)中的循环机制将该过程迭代T_c次后,获得每个节点的最终隐藏状态$\{h_1^{T_c}, h_2^{T_c}, \dots, h_K^{T_c}\}$,将初始隐藏状态和最终隐藏状态作为一个全连接网络层ϕ_c的输入来计算描述节点的输出特征,随后将该描述区域的特征信息送入基于LSTM的描述模型中以生成图像的描述语句,有

$$f_i^c = \phi_c(h_c^0, h_i^{T_c}). \tag{13}$$

为了使用LSTM生成描述语句,本文利用从前一个时刻提取到的描述区域特征初始化LSTM当前时刻的隐藏状态.

通过这种特征信息传递结构,使得不同语义层的特征从连接节点聚合相邻节点信息并传播自身节点信息到相邻节点的过程以实现各自特征的更新,迭代更新后的特征信息将被用于预测.

2.4　图像描述

本文将物体的名称属性经过词嵌入处理表示为名称属性向量a_i,将其和空间位置特征向量x_i分别作为高层语义和空间注意力机制的输入. 注意力机制根据LSTM网络的隐藏状态从描述区域的特征向量f_i^c中动态地选择与当前时刻生成单词有关的特征来指导单词的生成,选取目标物体所对应的空间位置特征向量$\{x_1, x_2, \dots, x_n\}$和名称属性向量$\{a_1, a_2, \dots, a_n\}$以及解码器中LSTM网络上一时刻的隐藏状态h^{t-1}来决定当前时刻选取的空间位置特征向量和名称属性向量的权重α_i^t. 有

$$e_i^t = f_{\text{att}}(x_i, a_i, h^{t-1}), \tag{14}$$

$$\alpha_i^t = \frac{\exp(e_i^t)}{\sum_{k=1}^{n}\exp(e_k^t)}. \tag{15}$$

其中:f_{att}是一个基于多层感知机的注意力机制模型,对其输出结果进行归一化得到目标物体所对应的特征在t时刻的权重分布,这些权重表示描述模型对图像中各个目标物体的关注程度,然后根据LSTM模型当前时刻的隐藏状态和前一时刻的输出计算得到当前时刻输出单词的概率分布. 模型以标识符⟨S⟩为标志开始生成描述语句,通过集束搜索的方式循环不断地生成单词,直至生成停止标识符⟨E⟩或者达到生成单词序列的最大长度.

3　实验和分析

3.1　数据集和设置

Visual Genome (VG)是当前最大的场景图像注释数据集,在实验中本文遵循与文献[18]相同的预处理和分割方法,训练集和测试集分别包含75 651幅和32 422幅图像,选择最常用的150个对象类和50个

关系类. COCO是研究图像描述问题的基准数据集,本文使用Karpathy分割方法[35]对其进行划分,使用113 287张图像和对应的人工标注描述作为训练集,分别选取5 000张图像和对应的人工标注描述作为验证集和测试集. 训练模型之前,先将人工标注描述中的所有单词转换成小写形式,标点符号使用空格替代,将出现频率超过5次的单词构成单词表,使用〈UNK〉替换出现频率小于5次的单词,以避免罕见单词不利于描述文本的生成.

使用在ImageNet上预先训练的具有VGG-16的Faster R-CNN检测器生成候选区域集合. 首先训练RPNs,使用批量为6、动量为0.9的SGD训练检测器,以初始学习率0.003联合训练整个模型. 在描述模型训练阶段,使用Adam[36]作为训练的优化器,学习率的初始值设置为0.000 2,批量设置为16. 使用的图像特征和物体特征的维度均设置为2 048,词向量的维度和长短期记忆网络中隐藏层的维度均设置为512,生成的描述文本最大长度设置为16.

总之,首先训练目标检测器,然后固定参数来训练场景图生成模型和图像描述模型,最后联合训练整个模型直到收敛.

3.2 评估指标

本文采用文献[18, 28]中使用的指标对所提出的模型进行评估: 1) 谓词分类:给定物体边界框,通过类标签从一组物体中预测给定物体对的关系标签; 2) 场景图分类:通过边界框预测物体的类标签,并预测每个物体对的关系标签; 3) 场景图生成:检测物体,并预测每对被检测物体之间的谓词; 4) 图像描述:根据图像的内容生成对应的描述文本.

由于不可能对图像中所有的关系进行注释,本文使用Top-K召回率(R@K)作为场景图生成的评估标准,使用图像描述常用的5个评价指标对描述结果进行测评.

3.3 与现有方法的比较

3.3.1 场景图生成

将本文提出的方法与目前几种先进的场景图生成方法在Visual Genome数据集上进行比较,实验结果如表1所示.

表1 在Visual Genome数据集上进行的测试结果

模型	场景图生成/%		场景图分类/%		谓词分类/%		平均值/%
	R@50	R@100	R@50	R@100	R@50	R@100	
VRD	0.3	0.5	11.8	14.1	27.9	35.0	14.9
IMP	3.4	4.2	21.7	24.4	44.8	53.1	25.3
MSDN	10.7	14.2	24.3	26.5	**67.0**	**71.0**	35.6
FREQ	23.5	27.6	32.4	34.0	59.9	64.1	40.2
Ours	**26.0**	**28.8**	**39.8**	**40.5**	51.3	56.7	**40.5**

表1给出了场景图生成的3个子任务的召回率以及召回率平均值. 由表1可知,本文提出的模型在场景图生成和分类任务上的性能优于所对比的模型,场景图生成任务的平均召回率也高于其他的对比模型. VRD[16]模型使用视觉特征和单词嵌入进行谓词分类,主要关注谓词分类性能. IMP[18]模型使用门循环单元进行信息的更新,主要针对局部关系上下文进行建模. 本文模型直接将不同语义层的特征连接并进行迭代更新,因此,模型的整体性能得到较大的提升. MSDN模型将VG数据集划分为关系和描述数据子集,将不同的视觉任务在对应的数据子集上进行训练,该模型通过引入描述分支来提供丰富的图像语义信息并用于谓词识别,因此,谓词分类任务的召回率会明显提升. FREQ[33]模型直接预测给定标签物体对的常见关系,因此,谓词分类任务得到了较高的召回率值. 本文模型使用不同语义层之间的特征信息进行迭代更新,以实现场景图生成,因此,在场景图生成和分类任务上的性能优于FREQ模型. 以上对比结果表明,本文所提出的模型同时从多个不同语义层进行特征信息的迭代更新生成场景图的准确性优于其他的场景图生成模型.

3.3.2 物体检测

物体检测识别作为视觉任务的基础,其准确度直接影响着高级视觉任务的生成. 本文使用Visual Genome作为物体检测的数据集,使用平均精度均值(mAP)和检测速度作为测量标准.

将本文所提出的检测方法与Faster R-CNN在相同数据集上进行测试,以检查增加图像描述和注意力机制是否可以提供有用的信息用于提升模型进行物体检测的准确性,对比结果如表2所示.

表 2　在 Visual Genome 数据集上进行的物体检测结果

物体检测	Faster R-CNN	ours-AM	ours
mAP/%	6.72	6.94	7.13
FPS 帧 /s	2.34	0.54	0.23

由表 2 可知:在增加具有注意力机制的图像描述情况下 (ours),模型的物体检测 mAP 比 Faster R-CNN 提高了 0.41 %;在增加没有注意力机制的图像描述情况下 (ours-AM),物体检测的 mAP 提升了 0.22 %. 这说明增加图像描述和注意力机制后,本文所提出的模型能够为物体检测识别提供额外有用的上下文信息,这些信息有助于提升物体检测识别的准确性. 本文所提出模型的结构相比于 Faster R-CNN 模型更加复杂,程序运行所需时间也相对变长,模型在增加具有融合注意力机制的图像描述情况下,检测速度为 0.23 帧/s.

3.3.3　图像描述

将本文提出的增加融合注意力机制的描述模型 (SA-Attention) 与当前最先进的描述模型在相同的 COCO 数据集上进行效果对比. 使用预先训练好的物体检测模型对注意力机制和解码器的参数进行训练,将训练好的模型作为基线模型 (Baseline).

本文使用 5 种常用的图像描述评价指标来检测所提出的融合注意力机制的描述效果,在指标上的得分越高,表示模型生成的描述效果越好. 表 3 给出了 SA-Attention 模型与几种对比模型的测试结果.

表 3　SA-Attention 模型与其他描述模型实验结果对比

模型	指标				
	B1	M	R	C	S
Baseline	0.741	0.257	0.541	1.019	0.191
Spatial-Attention[27]	0.752	0.263	0.552	1.060	0.196
Semantic-Attention[28]	0.709	0.243	0.543	1.042	–
Adaptive-Attention[29]	0.742	0.266	0.550	1.037	–
SA-Attention	**0.763**	**0.273**	**0.562**	**1.129**	**0.205**
SCA-CNN[25]	0.719	0.250	0.524	0.921	–
LSTM-A[37]	0.754	0.269	0.558	1.088	0.200
Up-Down[24]	0.772	0.270	0.564	1.135	0.203
SGAE[26]	**0.808**	**0.284**	**0.586**	**1.278**	**0.221**

表 3 中:BLEU-1(B1)[38] 指标通过分析生成的描述与参考描述中单词序列共同出现的程度来评价描述的准确度;METEOR(M)[39] 指标计算特定的序列匹配精度和召回率的调和平均值;ROUGE_L(R)[40] 指标通过度量生成的描述文本和参考描述文本之间最长公共子序列的相似度来评价生成描述的质量;CIDEr(C)[41] 指标通过计算生成的描述文本与参考描述文本之间的余弦相似度来衡量两者之间的相似性;SPICE(S)[42] 指标将生成的描述文本和参考描述文本解析为语法依赖树,随后使用基于规则的方法将依赖树映射成场景图以计算得分.

由表 3 可知:SA-Attention 模型在各个指标上的得分均高于基线模型和其他对比的注意力模型,这表明 SA-Attention 模型在增加融合注意力机制后能够有效提升图像描述的准确性. 与其他几种描述模型相比,SGAE 模型先从输入图像生成场景图,再将物体属性和场景图的结构等信息嵌入到描述模型,因此,其所生成的图像描述在上述指标上均具有很高的得分,描述性能优于 SA-Attention 模型. SA-Attention 模型在 CIDEr 指标上的得分分别比 Up-Down 和 SGAE 模型略低 0.006 和 0.149,在其他指标上的得分比 SGAE 模型略低 0.011～0.045. 除去 SGAE 模型之外,SA-Attention 模型在 METEOR 和 SPICE 指标上的得分也都高于其他对比模型,在 BLEU-1 和 ROUGE_L 指标上的得分也都接近于得分最高的 Up-Down 模型. 以上对比结果表明,本文提出的 SA-Attention 模型的描述性能优于其他对比的注意力模型,而且接近性能最优的描述模型,因此,SA-Attention 模型能够提取图像中物体的名称属性和空间位置特征信息,从而提升图像描述的准确性.

图 4 展示了 SA-Attention 模型、 基线模型和 Spatial-Attention 模型的部分图像描述效果,可以看出,基于 SA-Attention 模型所生成的图像描述结果最符合图像的实际内容.

图 4　使用不同模型生成图像描述效果示例

3.4　组件分析

为了进一步分析所提出模型中不同的设置对模型性能的影响,表 4 给出了不同的设置对场景图生成的 3 个子任务性能的影响.

表4 所提出模型在不同组件设置下的有效性

模型	信息传递	图像描述	注意力机制	场景图生成/%		场景图分类/%		谓词分类/%		平均值/%
				R@50	R@100	R@50	R@100	R@50	R@100	
1	–	–	–	15.7	18.5	23.6	26.9	37.2	44.8	27.8
2	√	–	–	23.8	26.5	36.7	37.4	46.9	52.4	37.3
3	√	√	–	25.6	27.3	38.6	39.7	49.4	54.8	39.2
4	√	√	√	**26.0**	**28.8**	**39.8**	**40.5**	**51.3**	**56.7**	**40.5**

对于特征信息传递,比较表4中的模型1和模型2可以得出,使用信息传递可以使模型的整体性能得到提升,模型的平均召回率提高了9.5%. 这表明在不同语义层之间进行的特征信息传递细化,不仅可以利用物体与关系短语之间的连接信息提升模型的性能,而且信息传递能够捕获有用的上下文信息以提升场景图生成任务的性能.

对于图像描述:比较表4中模型2和模型3可以得出,增加图像描述后,模型的平均召回率提高了1.9%;比较模型3和模型4可以得出,在图像描述中增加融合注意力机制来指导场景图生成的情况下,模型性能的平均召回率提高了1.3%. 这表明图像描述能够有效地指导场景图生成,提高图像描述的准确性有助于提升场景图生成的准确性. 因为增加融合注意力机制的图像描述模型可以生成更加符合图像实际内容的描述信息,同时也会为场景图生成提供额外的有用信息.

3.5 场景图像理解的可视化结果

场景图像理解的可视化结果如图5所示,本文将场景图像中的主要物体和关系进行检测,生成场景图并对场景图像中的显著区域(虚线框内)进行描述. 通过展示正确的物体检测、关系检测、场景图生成和图像描述来实现对场景图像的理解.

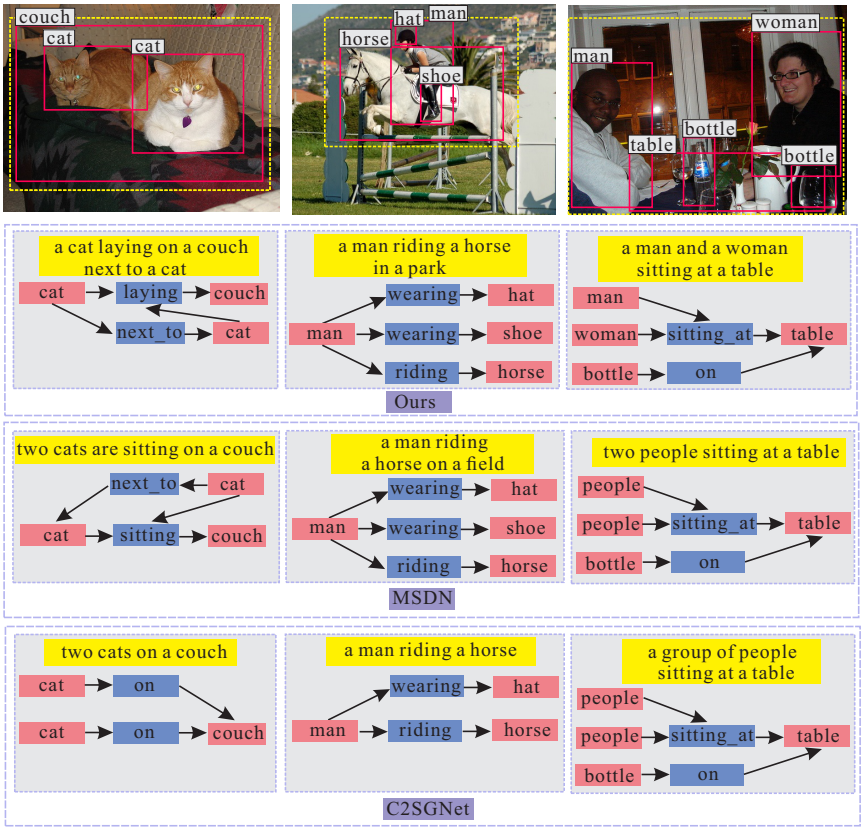


图5 场景图像理解的可视化结果

在图5中同时给出了本文所提出的模型与MSDN模型和C2SGNet模型实现场景图像理解的可视化结果. 对比图5中可视化结果的第1行和第2行可以得出,本文所提出模型生成的图像描述更符合图像的实际内容,描述准确性高于MSDN模型;对比图5中可视化结果的第1行和第3行可以得出,本文所提出模型在实现物体检测、场景图生成和图像描述这3个视觉任务的准确性都高于C2SGNet模型,这表明

本文所提出的基于多层语义特征的图像理解方法能够提升物体检测、场景图生成和图像描述的准确性,能够实现对场景图像更全面、更准确的理解.

4 结 论

本文提出了一种基于多层语义特征的图像理解方法,将场景理解中不同语义层的视觉任务同时进行求解,通过对物体检测、视觉关系检测和图像描述这3种不同语义层的视觉任务进行联合建模,实现了对场景图像更全面、准确的理解. 通过构建信息传递图,将不同语义层的特征信息同时进行迭代更新,使用更新后的语义特征预测物体和视觉关系以及生成场景图和图像描述,并引入融合注意力机制,提升了图像描述的准确性. 实验分析表明:本文提出的基于多层语义特征的图像理解方法能够有效利用图像中不同语义层的特征信息,提升物体检测、场景图生成和图像描述的准确性;通过将多个语义层的视觉任务进行联合求解,可以提升场景图像理解的整体效果.

参考文献(References)

[1] 李祥攀, 张彪, 孙凤池, 等. 基于多视角RGB-D图像帧数据融合的室内场景理解[J]. 计算机研究与发展, 2020, 57(6): 1218-1226.
(Li X P, Zhang B, Sun F C, et al. Indoor scene understanding by fusing multi-view RGB-D image frames[J]. Journal of Computer Research and Development, 2020, 57(6): 1218-1226.)

[2] Liu Y G, Yu J Z, Han Y H, et al. Understanding the effective receptive field in semantic image segmentation[J]. Multimedia Tools and Applications, 2018, 77(17): 22159-22171.

[3] Yatskar M, Zettlemoyer L, Farhadi A, et al. Situation recognition: Visual semantic role labeling for image understanding[C]. Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 5534-5542.

[4] Zitnick C L, Parikh D, Vanderwende L, et al. Learning the visual interpretation of sentences[C]. International Conference on Computer Vision. Sydney: IEEE, 2013: 1681-1688.

[5] Desai C, Ramanan D, Fowlkes C C, et al. Discriminative models for static human-object interactions[C]. Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 9-16.

[6] Yao B D, Li F F. Modeling mutual context of object and human pose in human-object interaction activities[C]. Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 17-24.

[7] Sadeghi M A, Farhadi A. Recognition using visual phrases[C]. Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 1745-1752.

[8] Li Y K, Ouyang W L, Zhou B L, et al. Scene graph generation from objects, phrases and region captions[C]. International Conference on Computer Vision. Venice: IEEE, 2017: 1270-1279.

[9] Shin D, Kim I. Deep image understanding using multilayered contexts[J]. Mathematical Problems in Engineering, 2018, 2018: 1-11.

[10] Krishna R, Zhu Y K, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[11] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]. European Conference on Computer Vision. Zurich: IEEE, 2014: 740-755.

[12] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580-587.

[13] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]. Advances in Neural Information Processing Systems. Montreal: IEEE, 2015: 91-99.

[14] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. 2020, arXiv: 2004.10934v1.

[15] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. European Conference on Computer Vision. Amsterdam: IEEE, 2016: 21-37.

[16] Lu C W, Krishna R, Bernstein M S, et al. Visual relationship detection with language priors[C]. European Conference on Computer Vision. Amsterdam: IEEE, 2016: 852-869.

[17] Li Y K, Ouyang W L. Wang X G, et al. ViP-CNN: Visual phrase guided convolutional neural network[C]. Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7244-7253.

[18] Xu D F, Zhu Y K, Choy C B, et al. Scene graph generation by iterative message passing[C]. Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3097-3106.

[19] Li S, Xiao T, Li H S, et al. Person search with natural language description[C]. Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5187-5196.

[20] Farhadi A, Hejrati M, Sadeghi M A, et al. Every picture tells a story: Generating sentences from images[C]. European Conference on Computer Vision. Berlin: IEEE, 2010: 15-29.

[21] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[22] Elliott D, Keller F. Image description using visual dependency representations[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: IEEE, 2013: 1292-1302.

[23] Verma Y, Gupta A, Mannem P, et al. Generating image descriptions using semantic similarities in the output space[C]. Computer Vision and Pattern Recognition. Portland: IEEE, 2013: 288-293.

[24] Anderson P, He X D, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]. Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 6077-6086.

[25] Chen L, Zhang H W, Xiao J, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning[C]. Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6298-6306.

[26] Yang X, Tang K H, Zhang H W, et al. Auto-encoding scene graphs for image captioning[C]. Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 10677-10686.

[27] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]. International Conference on Machine Learning. Lille: IEEE, 2015: 2048-2057.

[28] You Q Z, Jin H L, Wang Z W, et al. Image captioning with semantic attention[C]. Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4651-4659.

[29] Lu J S, Xiong C M, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]. Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3242-3250.

[30] Johnson J, Krishna R, Stark M, et al. Image retrieval using scene graphs[C]. Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3668-3678.

[31] Li Y K, Ouyang W L, Zhou B L, et al. Factorizable net: An efficient subgraph-based framework for scene graph generation[C]. European Conference on Computer Vision. Munich: IEEE, 2018: 346-363.

[32] Yang J W, Lu J S, Lee S, et al. Graph R-CNN for scene graph generation[C]. European Conference on Computer Vision. Munich: IEEE, 2018: 690-706.

[33] Zellers R, Yatskar M, Thomson S, et al. Neural motifs: Scene graph parsing with global context[C]. Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 5831-5840.

[34] Li Y, Tarlow D, Brockschmidt M, et al. Gated graph sequence neural networks[C]. International Conference on Learning Representations. San Juan: IEEE, 2016: 1-20.

[35] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[C]. Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3128-3137.

[36] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. 2014, arXiv: 1412.6980.

[37] Yao T, Pan Y W, Li Y H, et al. Boosting image captioning with attributes[C]. International Conference on Computer Vision. Venice: IEEE, 2017: 4904-4912.

[38] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation[C]. Annual Meeting on Association for Computational Linguistics. Philadelphia: IEEE, 2002: 311-318.

[39] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]. Transactions of the Association for Computational Linguistics. Michigan: IEEE, 2005: 65-72.

[40] Lin C Y, Hovy E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C]. North American Chapter of the Association for Computational Linguistics. Edmonton: IEEE, 2003: 71-78.

[41] Vedantam R, Zitnick C L, Parikh D, et al. CIDEr: Consensus-based image description evaluation[C]. Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 4566-4575.

[42] Anderson P, Fernando B, Johnson M, et al. Spice: Semantic propositional image caption evaluation[C]. European Conference on Computer Vision. Amsterdam: IEEE, 2016: 382-398.

作者简介

莫宏伟 (1973−) , 男, 教授, 博士生导师, 从事人工智能和类脑计算等研究, E-mail: honwei2004@126.com;

田朋 (1989−), 男, 博士生, 从事图像描述和场景理解的研究, E-mail: tianpeng0609@163.com.

(责任编辑：李君玲)