计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20190329
56(8):1721-1730,2019

一种残差置乱上下文信息的场景图生成方法

林 欣[1] 田 鑫[1] 季 怡[1] 徐云龙[2] 刘纯平[1,3]
[1](苏州大学计算机科学与技术学院 江苏苏州 215006)
[2](苏州大学应用技术学院 江苏苏州 215300)
[3](符号计算与知识工程教育部重点实验室(吉林大学) 长春 130012)
(xlin2017@stu.suda.edu.cn)

Scene Graph Generation Based on Shuffle Residual Context Information

Lin Xin[1], Tian Xin[1], Ji Yi[1], Xu Yunlong[2], and Liu Chunping[1,3]
[1](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)
[2](*Applied Technology College of Soochow University, Suzhou, Jiangsu* 215300)
[3](*Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun* 130012)

Abstract Scene graphs play an important role in visual understanding. Existing scene graph generation methods focus on the research of the subjects, the objects as well as the predicates between them. However, human being abstracts the relationships using spatial relation context, semantic context and interaction between scene objects for better understanding and reasoning as whole. In order to obtain the better global context representation and reduce the impact of dataset bias, we propose a new framework of scene graph generation, called as residual shuffle sequence model (RSSQ). Our method is made up of object decoding, residual shuffle and position embedding modules. Residual shuffle module is stacked with two basic structures including the random shuffle operation and the residual bidirectional LSTM. We implement the random shuffle on the hidden state of bidirectional LSTM by the process of iterative operation to reduce the impact of dataset bias, and extract the shared global context information by the residual connection structure. To strengthen the spatial relationship between pair-wise objects, the encoding is achieved using the relative position and area ratio of objects in position embedding module. The experimental results of three sub-tasks of different difficulty performed on Visual Genome dataset, demonstrate that the poposed method can generate better scene graphs under Recall@50 and Recall@100 settings due to better global context and spatial information.

Key words scene graph; visual relationship; context; residual bidirectional LSTM; object detection

摘 要 场景图在视觉理解中有着很重要的作用.现有的场景图生成方法对于主语、宾语以及主宾语间的视觉关系进行研究.但是,人类通过空间关系上下文、语义上下文和目标之间的互动信息来进行关系的理解和推理.为了获得更好的全局上下文表示,同时减少数据集偏差的影响,提出了一个新的场景图

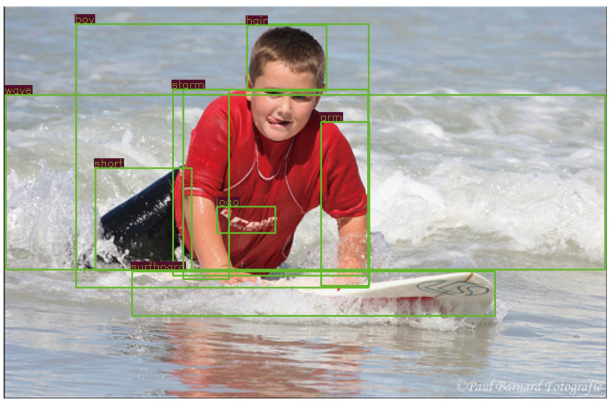
收稿日期:2019-06-03;**修回日期**:2019-06-20
基金项目:国家自然科学基金项目(61773272,61272258,61301299);吉林大学符号计算与知识工程教育部重点实验室项目(93K172016K08);江苏高校优势学科建设工程资助项目
This work was supported by the National Natural Science Foundation of China (61773272, 61272258, 61301299), the Program of the Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education (93K172016K08), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.
通信作者:刘纯平(cpliu@suda.edu.cn)

生成框架 RSSQ(residual shuffle sequence model).该框架由目标解码、残差置乱和位置嵌入 3 部分构成.残差置乱模块由随机置乱和残差连接的双向 LSTM 的基本结构叠加而成,利用迭代方式实现随机打乱双向 LSTM 的隐藏状态以减少数据集偏差影响,利用残差连接提取共享的全局上下文信息.在位置嵌入模块中,通过对目标的相对位置和面积比例的编码则可以增强目标对之间的空间关系.在数据集 Visual Genome 的 3 个不同层次子任务的实验中,证明了提出的 RSSQ 方法因全局上下文改善和空间关系增强,在 Recall@50 和 Recall@100 指标评价下,相对于现有方法能生成更好的场景图.

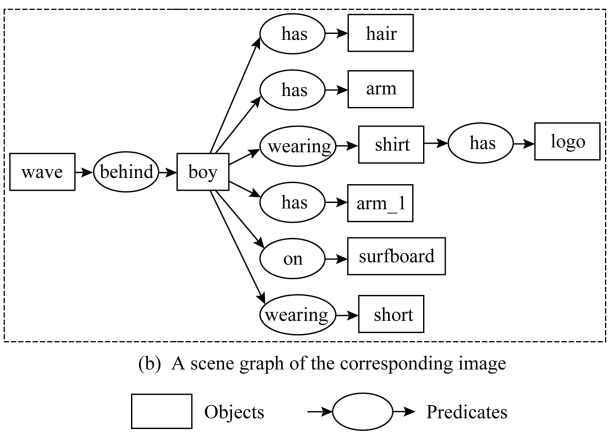
关键词 场景图;视觉关系;上下文;残差双向 LSTM;目标检测

中图法分类号 TP391

场景图[1]是真实图像中目标和目标间关系的精细化语义抽取,通过对预定义的目标实例、目标属性和目标对间关系进行预测来构建,常用三元组的结构化语言表示场景中目标间的交互.图 1 给出了一幅图像三元组关系表示的场景图实例,如〈boy-wearing-shirt〉.在场景图中,节点描述类别信息连同边界盒表示的目标实体,有向边则表示主、宾语间的关系类别.借助场景图对一幅图像可解释结构化表示的描述,图像被重构为连接图结构而不是孤立的目标实体,可以支持高层视觉智能任务,如图像检索[2]、目标检测[3-4]以及视觉问答[5-7]等视觉任务.由



(a) Image with bounding boxes



(b) A scene graph of the corresponding image

Fig. 1 A sample of a ground truth scene graph

图 1 场景图示意图

于手工标注海量图像的三元组关系描述格外昂贵,因此训练一个模型来自动生成高质量的场景图是近年来视觉理解的一种重要方向,再加上场景图表示需要推理复杂的依赖关系,高效准确地提取场景图也是一个极具挑战性的任务.

作为连接视觉与语言的桥梁,场景图生成任务是尽可能生成一个精确映射真实视觉场景的图表示.现有大多数基于目标的场景图方法,主要有基于目标检测和关系分类两阶段生成方法、基于目标和关系联合推理两大类.基于推理的场景图生成方法又可细分为基于消息传递[1,8-10]和全局上下文[11-12] 2 类.为得到更精准的目标标签,这类方法在候选场景图上进行消息传递与推理.

基于消息传递的方法中,首先提取目标区域的局部特征输入循环神经网络学习,其次使用相邻节点和边的表示来生成消息,并在图的拓扑结构中进行传递,最终获得主语、宾语和关系的最终表示结果.常见的消息传递策略包括迭代消息传递[1]、并行和串行消息传递[9]、空间加权消息传递[10]等.Xu 等人[1]最早提出基于迭代消息传递的场景图生成方法 IMP(iterative message passing).该方法首先通过 ROI-pooling[13]从 VGG-16 卷积层[14]中提取目标和关系的特征,然后将视觉特征分别输入节点和边 GRU(gated recurrent unit)[15]中,在之后的迭代过程中根据拓扑结构,利用相邻节点或边的隐藏状态生成消息,获取最终目标和关系表示.此外,还有一些改进的消息传递方法被提出,如并行和串行消息传递策略[9]可以更好地在目标和关系间传递信息;空间加权消息传递结构和空间敏感关系推理模块机制下的基于子图连接图[10]可有效加速推理过程和提高场景图生成效率.但是由于不完全的数据集标注,此类模型生成的消息受到局部上下文偏差的影响以及缺乏全局的视野.

基于视觉和语义特征候选场景图中节点间上下

文传递下更新节点和关系表示能更加有效地学习到可靠边的位置,减少不可能边的影响.NM(neural motifs)模型[11]是最具代表性的全局上下文方法,此外还有注意力图卷积网络[12]的场景图生成方法.相对于局部上下文方法局限于关系三元组进行消息传递,全局上下文方法在全图范围内进行上下文更新,从而获取更加全面的特征表示.在 NM 模型中,目标候选框的特征以一个固定的顺序被输入到双向 LSTM(long short-term memory)网络[16]中,从而获得图像的全局上下文,并通过连接主、宾语的全局上下文表示,实现对关系的分类.由于该类方法将原始图像中呈二维空间分布的目标排列成一个固定的从左至右的线性顺序,全局上下文信息受到破坏,使模型更倾向于学习到数据集的偏差,而不是真正的视觉关系表示,同时损失了空间信息,无法获得全面的全局上下文.

鉴于上述问题,本文以 NM 模型[11]为基础,提出了残差置乱上下文信息的场景图生成模型(residual shuffle sequence model, RSSQ),其主要贡献有 3 个方面:

1) 提出随机置乱策略,将固定顺序的隐藏状态迭代打乱重组.该策略就像纸牌游戏中的洗牌操作,可以加强目标节点和其他所有相邻节点的信息交换,提高模型的泛化能力,降低数据集偏差对场景图生成的影响.

2) 构建不同双向 LSTM 层之间的残差连接,获得不同层次的全局上下文信息,以形成更好的全局共享上下文表达,同时因残差的引入解决梯度消失问题.

3) 提出显式编码目标对间的位置信息嵌入,以增强场景图生成中的空间上下文,改善目标关系描述.

1　相关工作

场景图生成是近几年才发展起来的计算机视觉高级任务之一.与本文提出场景图生成方法密切相关联的工作主要有 NM 模型和残差连接.下面分别介绍这 2 个方面.

NM 模型[11]是一种代表性的全局上下文方法.该模型将场景图生成分为候选目标边界盒、区域标签和关系预测 3 个阶段.在候选目标边界盒预测阶段,计算边界盒区域内的上下文信息并进行传递;然后将全局上下文用于预测边界盒的标签,并基于全局上下文进行边预测;最后在融合上下文边界盒区域信息的基础上给边分配标签.具体实现中首先提取候选目标的局部特征,并以候选区域中心点在原图上的位置从左至右的线性顺序将局部特征输入双向 LSTM;然后用一个单向 LSTM 来解码目标类别,连同目标上下文输入到边上下文双向 LSTM 网络中;最后组合主、宾语特征,获取关系的最终表示.通过序列学习,NM 模型能够学到视觉场景的强规则化信息,但是具有复杂空间分布和丰富语义信息的图像被抽象为一个固定次序线性序列的简单操作造成了重要信息损失,如场景中的空间位置信息丢失;再加上双向 LSTM 的强记忆能力使得 NM 模型更容易学习到数据集的偏差.

与本文提出场景图生成方法相关的另一个工作是残差连接.残差连接的关键思想是在网络层之间增加短路连接,提供额外的梯度路径[17].通过残差连接,非常深的卷积网络[17]被应用与图像分类和检测.残差连接在深层卷积神经网络中的应用,提高了模型的泛化能力,解决了模型的"退化"问题.最近,Kim 等人[18]提出了在 LSTM 模型中增加残差连接的方法,并将该方法应用于远场语音识别,证明了残差连接可以提供短路,解决梯度消失问题.鉴于深度学习中,不同的网络层可以表示低/中/高不同层次的特征[19],因此,在不同层次的 LSTM 中建立残差连接能够更好地学习抽象视觉关系,减少梯度消失问题.NM 模型在双向 LSTM 中使用高速连接的设计,在时间维度上解决了梯度消失问题,但是随着层数的增加,建立了高速连接的 LSTM 仍然存在退化问题[20],同时在空间维度上高速连接使得训练过程更加困难,残差连接解决了这个问题[18].

2　RSSQ 方法

为了获取更优的关系表示以生成更精确的场景图,提出了 RSSQ 方法.该方法主要由目标解码模块、残差置乱模块以及位置嵌入模块 3 个部分组成,其整体框架如图 2 所示.为了简洁和方便,下文双向 LSTM 隐藏状态均表述为上下文信息.

场景图中的视觉关系包括目标和谓词.对于目标的提取,利用 Faster RCNN 模型[21]给出初始目标分类预测 $\boldsymbol{o}_i$,然后在目标解码模块中,利用初始目标 $\boldsymbol{o}_i$ 解码上下文信息 $\boldsymbol{h}_{i,d}$ 以进一步分类目标获得场景图中目标 $\hat{\boldsymbol{o}}_i$ 的表示:

$$\hat{\boldsymbol{o}}_i = \arg\max(fc(\boldsymbol{h}_{i,\mathrm{d}})), \tag{1}$$

其中,$fc(\cdot)$表示全连接,d 表示目标解码模块.主语

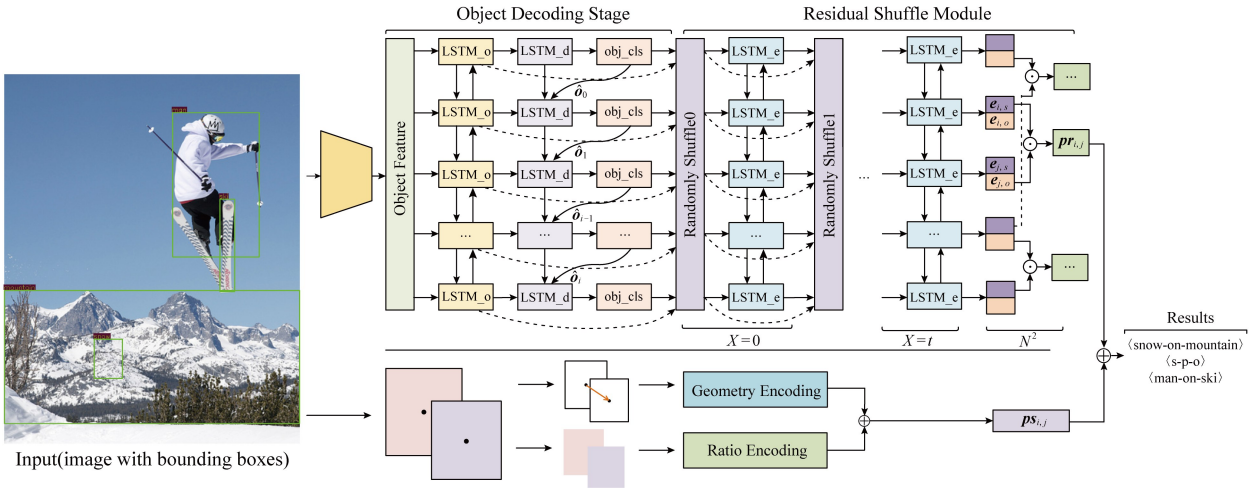


Fig. 2 The framework of our Residual Shuffle Sequence Model (RSSQ)

图 2 残差置乱上下文信息场景图生成方法框架

目标 i 和宾语目标 j 之间的谓词表示由置乱残差边上下文表示 $\boldsymbol{pr}_{i,j}$ 以及位置嵌入向量 $\boldsymbol{ps}_{i,j}$ 的最大全连接获得.谓词表示为

$$rel_{i,j}=\arg\max(fc(\boldsymbol{pr}_{i,j},\boldsymbol{ps}_{i,j})). \tag{2}$$

2.1 目标解码

目标解码阶段的主要目的是实现目标分类.该模块首先使用 Faster RCNN[21] 来进行目标的预分类以及目标边界盒的回归.由于 Faster RCNN 中,目标分类是不考虑上下文信息的.为了引入上下文信息,采用 NM 模型[11] 中的目标上下文模块构建目标预测的上下文表示.

目标上下文信息 $\boldsymbol{h}_{i,\mathrm{o}}$ 提取是利用中心点偏移从左至右将其目标特征向量 f_i 输入到高速双向 LSTM[16] 中获得,即:

$$\boldsymbol{h}_{i,\mathrm{o}}=\mathrm{biLSTM}(\boldsymbol{f}_i). \tag{3}$$

目标的分类向量由目标上下文信息 $\boldsymbol{h}_{i,\mathrm{o}}$ 输入目标解码 LSTM 获得,即:

$$\boldsymbol{h}_{i,\mathrm{d}}=\mathrm{LSTM}(\boldsymbol{h}_{i,\mathrm{o}}). \tag{4}$$

2.2 残差置乱

置乱操作被定义成一个将固定有序序列转换成随机序列的映射函数.映射函数采用随机数函数.假设目标序列初始排列顺序为 $\boldsymbol{I}=(1,2,\cdots,i,\cdots,N)$,$N$ 为图像中目标总数,i 为第 i 个目标.经过 t 轮置乱,新的顺序被更新为序列 $\hat{\boldsymbol{I}}^{t-1}$,其表达式为

$$\hat{\boldsymbol{I}}^{t-1}=(1,2,\cdots,\hat{i}^{t-1},\cdots,N)=\mathrm{shuffle}^{t-1}(\boldsymbol{I}), \tag{5}$$

其中,$\hat{i}^{t-1}$ 表示第 $t-1$ 次第 i 个目标,$t\in\{1,2,\cdots,M\}$.

残差置乱模块的输入由目标上下文编码的隐藏状态和词向量编码 2 部分拼接而成:

$$\boldsymbol{c}^0_{\hat{i}^0}=\mathrm{concat}(\boldsymbol{h}_{\hat{i}^0,\mathrm{o}},\boldsymbol{emb}_{\hat{i}^0}), \tag{6}$$

其中,$\boldsymbol{emb}_{\hat{i}^0}$ 是目标 $\hat{i}^0$ 的语义词向量,经过 $t+1$ 次置乱,边上下文输入双向 LSTM 的隐藏状态为

$$\boldsymbol{h}^t_{\hat{i}^t}=\mathrm{biLSTM}(\boldsymbol{c}^t_{\hat{i}^t}). \tag{7}$$

隐藏状态 $\boldsymbol{h}^t_{\hat{i}^t}$ 经置乱之后,新的隐藏状态序列则为 $\boldsymbol{h}^t_{\hat{i}^{t+1}}$.由于双向 LSTM 层间采用残差连接方式,所以 $t+1$ 轮置乱迭代之后,残差的边上下文信息 $\boldsymbol{c}^t_{\hat{i}^t}$ 更新为

$$\boldsymbol{c}^t_{\hat{i}^t}=\mathrm{concat}(\boldsymbol{h}^{t-1}_{\hat{i}^t},\boldsymbol{h}^{t-2}_{\hat{i}^t}),t\geqslant 2. \tag{8}$$

特别地,当 $t=1$ 时,$\boldsymbol{c}^1_{\hat{i}^1}=\mathrm{concat}(\boldsymbol{h}^0_{\hat{i}^1},\boldsymbol{c}^0_{\hat{i}^1})$.经过 $t+1$ 轮置乱后,利用全连接层将提取的边上下文信息 $\boldsymbol{h}^t_{\hat{i}^t}$ 分解成主宾成分:

$$\boldsymbol{e}_{\hat{i}^t,\mathrm{s}},\boldsymbol{e}_{\hat{i}^t,\mathrm{o}}=fc(\boldsymbol{h}^t_{\hat{i}^t}). \tag{9}$$

其中,$\boldsymbol{e}_{\hat{i}^t,\mathrm{s}},\boldsymbol{e}_{\hat{i}^t,\mathrm{o}}$ 分别表示目标对的主语和宾语成分,每对可能的目标对主语 $\hat{i}^t$ 和宾语 $\hat{j}^t$,对应原始编号 i,j.

最终残差边上下文表示 $\boldsymbol{pr}_{i,j}$ 为

$$\boldsymbol{pr}_{i,j}=\boldsymbol{e}_{\hat{i}^t,\mathrm{s}}\odot\boldsymbol{e}_{\hat{j}^t,\mathrm{o}}. \tag{10}$$

其中,$\odot$ 表示点乘运算.

2.3 位置嵌入

给定主语包围盒 $\boldsymbol{box}_i=(x_i,y_i,w_i,h_i)$,宾语包围盒 $\boldsymbol{box}_j=(x_j,y_j,w_j,h_j)$,主宾语间的相对几何特征 $\boldsymbol{PE}$ 和区域比特征 $\boldsymbol{A}_{\mathrm{up}}$,位置嵌入特征 $\boldsymbol{ps}_{i,j}$ 则可通过一个全连接层的融合得到:

$$\boldsymbol{ps}_{i,j}=fc(\boldsymbol{PE},\boldsymbol{A}_{\mathrm{up}}). \tag{11}$$

主、宾语间的相对几何特征 $\boldsymbol{PE}$ 是一个高维嵌入表示.为了获取平移和尺度不变的相对几何特征,对主宾语间的 4 维相对几何特征进行对数转换,转换后的相对几何特征为

$$\boldsymbol{pos}=\left(\log\left(\frac{|x_i-x_j|}{w_i}\right),\log\left(\frac{|y_i-y_j|}{h_i}\right),\log\left(\frac{w_i}{w_j}\right),\log\left(\frac{h_i}{h_j}\right)\right). \tag{12}$$

在本文实验中,根据文献[22]的方法,通过正弦和余弦函数分别计算主、宾语间的相对几何特征 $\boldsymbol{PE}$ 的奇数($2m+1$)和偶数($2m$)维度的变换特征,将4维相对几何特征 $\boldsymbol{pos}$ 换为64维表示.变换公式分别为

$$\boldsymbol{PE}_{(pos,2m)}=\sin(\boldsymbol{pos}/1000^{2m/d_{\mathrm{model}}}), \tag{13}$$

$$\boldsymbol{PE}_{(pos,2m+1)}=\cos(\boldsymbol{pos}/1000^{2m+1/d_{\mathrm{model}}}). \tag{14}$$

除了相对几何位置关系,目标对间的空间关系通过目标对之间面积关系和重叠关系来进一步增强[23].文献[23]中,通过相对位置、面积、形状等描述空间分布.受到该文献启发,本文引入4维区域比特征 $\boldsymbol{A}_{i,j}$,并利用一个ReLu函数激活的全连接层将其转换至64维:

$$\boldsymbol{A}_{\mathrm{up}}=\mathrm{ReLu}(fc(\boldsymbol{A}_{i,j})). \tag{15}$$

区域比特征 $\boldsymbol{A}_{i,j}=(V_{i,j},V_{o,i},V_{o,j},V_{o,u})$ 由1个面积比 $V_{i,j}$ 和3个重叠比 $V_{o,i},V_{o,j},V_{o,u}$ 构成:

$$\begin{aligned}V_{i,j}&=\frac{A(b_i)}{A(b_j)},\\ V_{o,i}&=\frac{A(o_{i,j})}{A(b_i)},\\ V_{o,j}&=\frac{A(o_{i,j})}{A(b_j)},\\ V_{o,u}&=\frac{A(o_{i,j})}{A(u_{i,j})}.\end{aligned} \tag{16}$$

其中,$A(b_i)$ 表示包围盒 box_i 的面积,$A(o_{i,j})$ 表示包围盒的重叠面积,$A(u_{i,j})$ 表示主宾语的外包围盒面积.

3　实验与结果分析

实验在公开数据集Visual Genome(VG)[24]上展开.为了验证提出RSSQ方法场景图生成性能,进行了模型本身的消融分析,同时进一步在关系分类、场景图分类和场景图生成3个不同层次子任务上进行方法性能的评价.

3.1　数据集及评价指标

Visual Genome数据集是一个人工标注的视觉关系数据集.根据不同的数据预处理方式和数据划分方法,存在多种不同的版本[8,11-12,25].在实验中,使用最普遍使用的数据预处理和数据集划分方法[1],其中训练集和测试集分别有75 651图像和32 422图像.保留了最常见的150类目标以及50类关系,每张图像平均有11.5个目标和6.2个关系.

场景图生成任务的目的是定位预定义的目标以及预测目标对间的关系.整个任务被分成3个子任务:

1) 关系分类任务(predicate classification, PredCls).给定真实目标框以及真实标签,需要预测目标对间关系;

2) 场景图分类任务(scene graph classification, SGCls).给定真实的目标边界盒,需要预测目标标签和目标对间关系;

3) 场景图生成任务(scene graph generation, SGGen).给定一张图像,需要检测其中的目标和关系.

实验评价指标采用Recall@K,缩写为R@K,是置信度最高的 K 个分类结果在关系真值中所占比例.本文根据在Visual Genome数据集中证明结论:随机生成一个三元组关系Recall@100约为0.000 089[24],在实验中将 K 取值为50和100.

3.2　RSSQ方法整体定量分析

实验中,以场景图中3个子任务为目标,将RSSQ方法与一些现存模型进行对比,包括Language Priors(LP)模型[26]、IMP模型[1]、Graph R-CNN(GR)模型[12]以及NM模型[11].实验结果如表1所示:

Table 1　Comparison with Some Existing Works

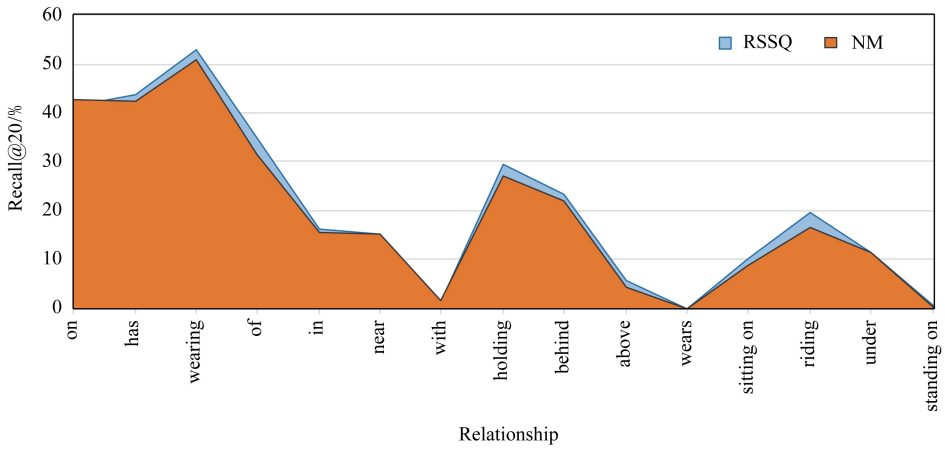
表1　RSSQ方法与现有方法对比实验结果

Methods	PredCls		SGCls		SGGen		SGGen*	
	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
PR[26]	27.9	35.0	11.8	11.4	0.3	0.4		
IMP[1]	44.8	53.0	21.7	24.4	3.4	4.2		
GR[12]	54.2	59.1	29.6	31.6	11.4	13.7		
NM[11]	65.2	67.1	35.8	36.5			27.2	30.3
RSSQ(ours)	**65.7**	**67.5**	**36.7**	**37.4**	**20.7**	**25.7**	**27.4**	**30.6**

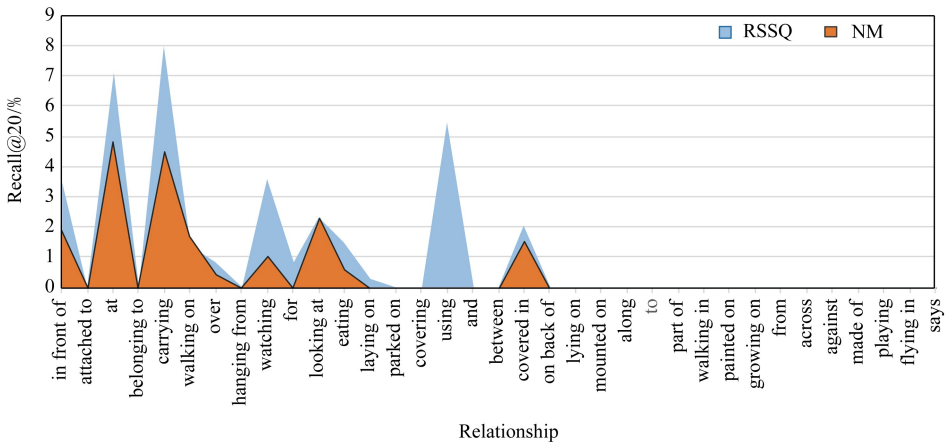
IMP 模型[1]主要针对局部关系上下文进行建模,丢失了全局上下文的视野.GR 模型[12]使用特定线性变换方法根据相邻节点进行节点表示更新,但是更新的策略相对简单.NM 模型[11]通过双向 LSTM 网络生成边上下文,丢失了结构化信息.从表 1 中可以看出,提出的 RSSQ 方法在 3 个子任务中都超过了现有方法.相对于 2018 年 CVPR 的 NM 模型,在子任务 SGCls 上超过 0.9%,在 PredCls 子任务上超过 0.5%.在 SGGen 子任务上,提出方法超过 GR 模型 12%.这表明提出 RSSQ 方法可以更加有效地生成场景图.

为了更进一步精确地对比提出地 RSSQ 方法和 NM 模型在分类性能上的改进.图 3 给出了在 SGCls 子任务中 Recall@20 设置上进行的关系分类准确率统计分析.横坐标上关系类别以出现频率的降序排列,只有在关系三元组全部被预测正确,包括主宾语和关系,才会被统计.图 3 给出了根据频率将关系分为高频(a)、中低频(b)2 个部分区段的实验对比.在高频段(图 3(a)),NM 模型和 RSSQ 方法对关系频率高的分类均表现良好,在部分关系类别中,提出的 RSSQ 方法相对于 NM 模型有微弱提升.

在中频区域(如图 3(b)所示),NM 模型的分类准确率较低,这是因为 NM 模型学到更多的数据集偏差而并非真正理解关系.提出的 RSSQ 方法在这个区间的关系分类精度有相对大的提升,比如 of, holding, behind, above, riding, at, carrying, using 以及 covered in 关系类别.受益于更好的全局上下文特征,提出的 RSSQ 方法在抽象关系分类精度方面有较明显提升,如 holding(+2.36%)、riding(+4.76%)、carrying(+9.75%)以及 using(+6.79).基于位置嵌入对位置信息的增强,提出的 RSSQ 方法对位置关系分类精度也有较大提升,如 of(+2.43%)、behind(+1.12%)、above(+1.55%)、at(+2.14%)以及 covered in(+2.55%).在低频段的分类识别,2 个模型



(a) Recall@20 of the relationship categories of high frequency



(b) Recall@20 of the relationship categories of medium and low frequency

Fig. 3 The accuracy of each relationship categories of SGCls of R@20 setting

图 3 关系分类逐类分析

均没什么表现,这就需要更多研究,比如少量学习[27].

总之,由于 Visual Genome 是一个严重不均衡的数据集,使大多模型更容易学习数据集偏差.提出的 RSSQ 方法在中等频率区间性能的明显提升,表明提出的 RSSQ 方法更少地受数据集偏差的影响,在一定程度上较好地改善了数据偏差对关系分类的影响.

3.3　残差置乱模块评价

基于 NM 模型[10]中 4 层 LSTM 层组成的边上下文模块(如图 4(a)所示),本文通过置乱模块和残差连接基本架构单元来构成残差置乱模块.通过对图 4(a)分别插入 1,2,4 次置乱层和残差连接构成 3 种残差置乱模块结构 e1,e2 和 e4,如图 4(b)~4(d)所示.

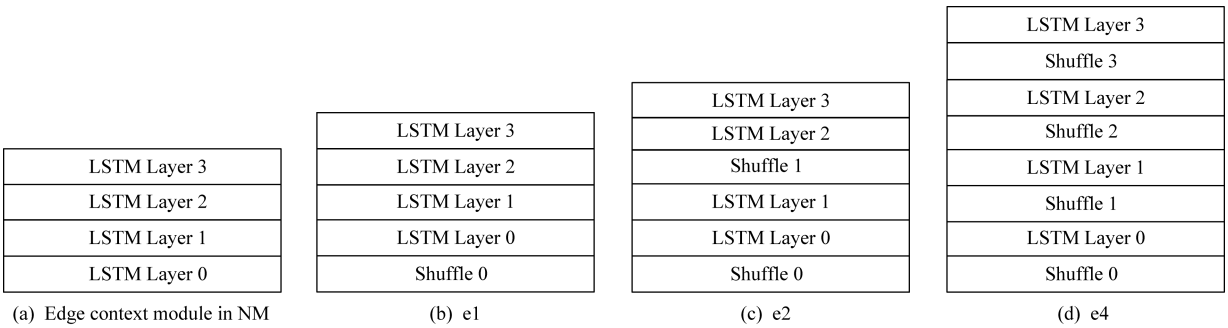


Fig. 4　The initial edge context module in NM[10] and structures of residual shuffle module insertion

图 4　残差置乱模块示意图

由于 NM 模型[10]没有给出未经微调的 SGGen 子任务的实验结果,残差置换模块的实验分析在 PredCls 和 SGCls 两个子任务上进行.此外,也进行了 LSTM 层之间的原始设置以及残差连接 2 种不同连接方式的实验.如表 2 所示,通过置乱操作,在 SGCls 任务中有 0.3%相对提升;通过残差连接,在 PredCls 子任务和 SGCls 子任务分别有 0.5%和 0.7%的相对提升.在单纯加入置乱操作的设置中,PredCls 子任务中有些许性能下降,这是由于 PredCls 使用目标标签真值,置乱破坏了关系的固定模式.从实验结果来看,置乱操作不断地打乱目标序列输入次序,在训练迭代过程中,即使是同一条训练数据也会有不同的输入次序,增加了模型的鲁棒性,提高了模型的泛化能力.残差连接融合了不同层次的边上下文,在不同 LSTM 层间建立短路,从而减少梯度消失问题,获取了更丰富语义的边上下文.

Table 2　Evaluation of the Residual Shuffle Module

表 2　残差置乱模块分析

Sub-tasks	Metrics	NM[10]	e1	e2	e4	e2r	e4r
Connections			raw	raw	raw	res	res
PredCls	R@50	65.2	64.96	65.07	64.87	65.38	**65.67**
	R@100	67.1	66.79	66.94	66.79	67.18	**67.47**
SGCls	R@50	35.8	36.04	36.11	35.98	36.39	**36.47**
	R@100	36.5	36.76	36.83	36.74	37.12	**37.17**

Note: "raw" means regular connection of LSTM layers, and "res" means residual connection.

3.4　消融实验

为进一步分析提出的 RSSQ 方法中残差置乱和位置嵌入 2 个模块对场景图生成的性能影响,表 3 给出了在 3 个子任务上的消融学习结果.这部分实验以 NM 模型为基准模型,单纯用残差置乱模块替换 NM 模型中的边上下文提取模块,在 PredCls 子任务和 SGCls 子任务中分别有 0.5%和 0.7% 的提升.单纯将位置嵌入模块添加到 NM 模型的边上下文模块中,在 PredCls 子任务和 SGCls 子任务中有些许提升.在 SGGen 子任务的实验中,位置嵌入模块与 NM 模型的结合是残差置换与 NM 模型结合,是提出 RSSQ 方法中性能表现最好的组合.提出的 RSSQ 方法在 2 个子任务 PredCls 和 SGCls 是表现最好的.综上分析,残差置乱和位置嵌入 2 个模块部分缓解了数据集偏差和全局上下文共享问题,完整的 RSSQ 方法在 3 个子任务中的综合表现良好.

Table 3 Ablation Study

表 3 消融实验结果

Methods	Residual Shuffle Module	Spatial Embedding	PredCls		SGCls		SGGen	
			R@50	R@100	R@50	R@100	R@50	R@100
NM[11]	Exclude	Exclude	65.2	67.1	35.8	36.5		
1	Include	Exclude	65.67	67.47	36.47	37.16	20.50	25.50
2	Exclude	Include	65.18	66.98	35.92	36.61	**20.74**	**25.91**
3(RSSQ)	Include	Include	**65.72**	**67.48**	**36.67**	**37.38**	20.64	25.66

3.5 部分场景图可视化结果

为了更直观展示提出的 RSSQ 方法在场景图生成的效果,图 5、图 6 给出了场景图可视化结果.其中图像中给出的是真值标签的边界盒,场景图给出了 SGCls 子任务中生成场景图和真值场景图的对比,方框表示目标实体,有向箭头从主语指向宾语,

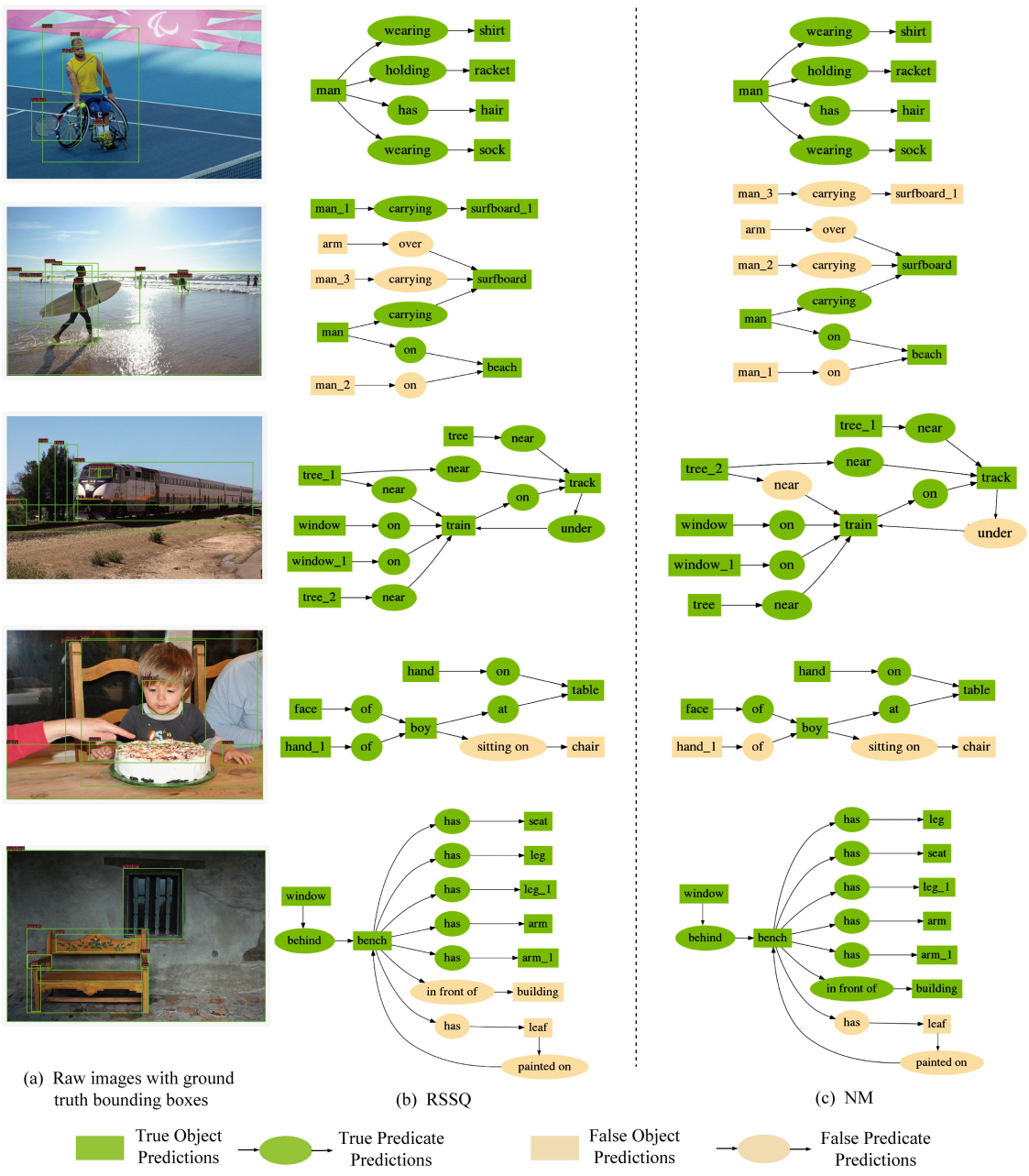


Fig. 5 Qualitative results of SGCls

图 5 场景图分类结果可视化结果

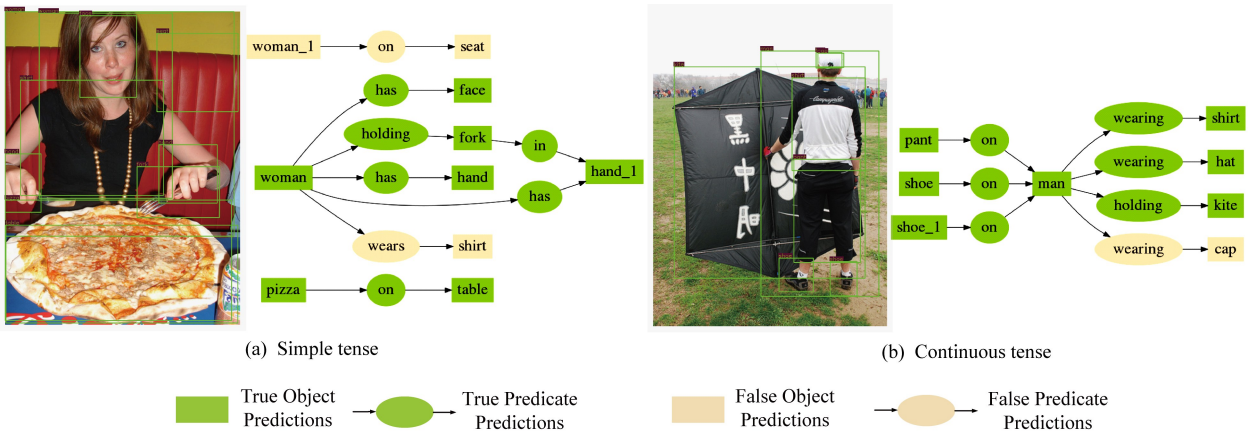


Fig. 6　Errors caused by tense disagreements

图 6　时态不一致引起的错误示例

椭圆形表示关系.每个给出的具体样例中的完整场景图是真值描述的场景图,其中深色底纹表示正确预测,浅色底纹表示错误预测.图 5(a)是原始带有目标真值标签的原始图像,图 5(b)给出的是 RSSQ 方法生成的场景图,图 5(c)是 NM 模型[10]生成的场景图.

图 6 给出了由于谓词的时态不一致性带来的关系分类错误,如图 6(a)中 wears 和图 6(b)中的 wearing.从图 5 第 1 行样例可以看出,RSSQ 方法和 NM 模型[10]均能比较吻合地生成比较简单的场景图.从图 5 第 3 行与第 5 行样例可以看出,RSSQ 方法相对于 NM 模型[10]改进了相对位置关系(near,under,in front of)的分类.从图 5 第 2 行与第 5 行样例可以看出,RSSQ 方法在中频区间的关系类别(carrying,in front of)有一定改进,缓解了数据集偏差问题.图 5 第 4 行样例说明,RSSQ 方法对于高频区间的关系分类(如 of)也有改进.

4　总　　结

鉴于场景图生成方法更多的学习数据集偏差,本文从残差置乱和位置嵌入角度改进 NM 模型,提出了一个新的基于残差置乱上下文信息的场景图生成方法(RSSQ).置乱策略有效地改善了数据集偏差对场景图生成的影响,尤其是在中频段的关系分类性能的提升比较明显;残差连接在不同 LSTM 层之间建立短路连接,完成不同层次的信息交换,较好解决了全局上下文信息共享,此外,残差连接还解决了梯度消失问题.位置嵌入从面积比和重叠比角度整合目标位置信息,也有效地提升了提出的 RSSQ 方法对位置关系分类的性能.在 Visual Genome 数据集的实验中验证了提出的 RSSQ 方法可行且高效,可以更少地受到数据集偏差的影响.

参　考　文　献

[1] Xu Danfei, Zhu Yuke, Choy C B, et al. Scene graph generation by iterative message passing [C] //Proc of the 2017 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3097-3106

[2] Johnson J, Krishna R, Stark M, et al. Image retrieval using scene graphs [C] //Proc of the 2015 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3668-3678

[3] Sadeghi M A, Farhadi A. Recognition using visual phrases [C] //Proc of the 2011 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 1745-1752

[4] Hu Han, Gu Jianyuan, Zhang Zheng, et al. Relation Networks for Object Detection [C] //Proc of the 2018 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 3588-3597

[5] Teney D, Liu Lingqiao, Den Hengel A V, et al. Graph-structured representations for visual question answering [C] //Proc of the 2017 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3233-3241

[6] Tang Kaihua, Zhang Hanwang, Wu Baoyuan, et al. Learning to compose dynamic tree structures for visual contexts [J]. arXiv preprint, arXiv:1812.01880v1, 2018

[7] Yu Jun, Wang Liang, Yu Zhou. Research on visual question answering techniques [J]. Journal of Computer Research and Development, 2018, 55(9): 1946-1958 (in Chinese)
(俞俊, 汪亮, 余宙. 视觉问答技术研究[J]. 计算机研究与发展, 2018, 55(9): 1946-1958)

[8] Li Yikang, Ouyang Wanli, Zhou Bolei, et al. Scene graph generation from objects, phrases and region captions [C]// Proc of the 2017 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1270-1279

[9] Li Yikang, Ouyang Wanli, Wang Xiaogang, et al. ViP-CNN: Visual phrase guided convolutional neural network [C] //Proc of the 2017 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7244-7253

[10] Li Yikang, Ouyang Wanli, Zhou Bolei, et al. Factorizable Net: An efficient subgraph-based framework for scene graph generation [C] //Proc of the 2018 IEEE European Conf on Computer Vision. Piscataway, NJ: IEEE, 2018: 346-363

[11] Zellers R, Yatskar M, Thomson S, et al. Neural motifs: Scene graph parsing with global context [C] //Proc of the 2018 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 5831-5840

[12] Yang Jianwei, Lu Jiasen, Lee S, et al. Graph R-CNN for scene graph generation [C] //Proc of the 2018 IEEE European Conf on Computer Vision. Piscataway, NJ: IEEE, 2018: 690-706

[13] Girshick R B. Fast R-CNN [C] //Proc of the 2015 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1440-1448

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C] //Proc of the 2015 IEEE Int Conf on Learning Representations. Piscataway, NJ: IEEE, 2015

[15] Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint, arXiv: 1412.3555, 2014

[16] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780

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

[18] Kim J, Elkhamy M, Lee J, et al. Residual LSTM: Design of a deep recurrent architecture for distant speech recognition [C] //Proc of the 2017 Conf of the Int Speech Communication Association, 2017: 1591-1595. https://www.isca-speech.org/archive/Interspeech_2017/pdfs/0477.PDF

[19] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C] //Proc of the 2014 IEEE European Conf on Computer Vision. Piscataway, NJ: IEEE, 2014: 818-833

[20] Zhang Yu, Chen Guoguo, Yu Dong, et al. Highway long short-term memory RNNS for distant speech recognition [C] //Proc of the 41st Int Conf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2016: 5755-5759

[21] Ren S, 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 and Machine Intelligence, 2017, 39(6): 1137-1149

[22] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the Conf and Workshop on Neural Information Processing Systems. New York: Curran Associates, 2017: 5998-6008

[23] Zhu Yaohui, Jiang Shuqiang, Li Xiangyang, et al. Visual relationship detection with object spatial distribution [C] //Proc of the 2017 IEEE Int Conf on Multimedia and Expo, Piscataway, NJ: IEEE, 2017: 379-384

[24] Krishna R, Zhu Yuke, 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

[25] Dai Bo, Zhang Yuqi, Lin Dahua, et al. Detecting visual relationships with deep relational networks [C] //Proc of the 2017 IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 3298-3308

[26] Lu Cewu, Krishna R, Bernstein M S, et al. Visual relationship detection with language priors [C] //Proc of the 2016 IEEE European Conf on Computer Vision. Piscataway, NJ: IEEE, 2016: 852-869

[27] Li Zhenguo, Zhou Fengwei, Chen Fei, et al. Meta-SGD: Learning to learn quickly for few shot learning [J]. arXiv preprint, arXiv:1707.09835, 2017

Lin Xin, born in 1995. MSc candidate. Student member of CCF. Her main research interests include computer vision and scene understanding.

Tian Xin, born in 1996. MSc candidate. His main research interests include computer vision and image processing.

Ji Yi, born in 1973. PhD, associate professor. Member of CCF. Her main research interests include 3D action recognition and complex scene understanding.

Xu Yunlong, born in 1964. BSc, associate professor. His main research interests include reinforcement learning, natural language processing, operating system and big data.

Liu Chunping, born in 1971. PhD, professor, PhD supervisor. Her main research interests include computer vision, image analysis and recognition, in particular in domains of visual saliency detection, object detection and scene understanding.