基于标签层次结构的视觉关系检测模型

王元龙[1],雷 鸣[1],王智强[1],张 虎[1],李 茹[1,2],梁吉业[1,2]
(1. 山西大学计算机与信息技术学院,山西太原 030006;
2. 山西大学计算智能与中文信息处理教育部重点实验室,山西太原 030006)

摘 要: 视觉关系检测是在目标识别的基础上,进一步检测出目标之间的关系,属于视觉理解和推理的关键技术. 然而,由于关系标签视觉上的相似性以及数据不平衡问题造成少样本的尾部关系检测召回率较低. 为了提高尾部关系的检测效果,本文将关系标签进行粗细粒度划分构建了标签的层次结构表示,提出了基于标签层次结构的视觉关系检测模型. 模型利用视觉关系之间的相似性以及数据带有的偏见性构建关系标签的层次结构表示,以此将关系区分为粗粒度关系和细粒度关系,使尾部关系在由粗粒度到细粒度的结构上获得更多的关注. 同时,针对标签层次结构的性质设计其损失函数,该损失函数通过结构化信息逐层学习不同类别关系之间的差异,使模型更好的检测尾部细粒度关系. 分别在公开数据集 Visual Relationship Detection(VRD)和 Visual Genome(VG)中验证了本文模型检测尾部关系的效果. 与现有模型相比,在 VRD 数据集中平均召回率 mR@20、mR@50 和 mR@100 分别提高了 0.62%、1.57% 和 2.47%;在 VG 数据集中,mR@20、mR@50 和 mR@100 分别提高了 0.67%、0.83% 和 1.15%.

关键词: 视觉关系检测;标签层次结构表示;长尾分布;粗粒度关系;细粒度关系
基金项目: 国家重点研发计划(No.2020AAA0106100);国家自然科学基金(No.62176145)
中图分类号: TP391.7 **文献标识码:** A **文章编号:** 0372-2112(2023)12-3496-11
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221050

Visual Relationship Detection Model Based on Label Hierarchy

WANG Yuan-long[1], LEI Ming[1], WANG Zhi-qiang[1], ZHANG Hu[1], LI Ru[1,2], LIANG Ji-ye[1,2]
(1. *School of Computer Science and Technology, Shanxi University, Taiyuan, Shanxi* 030006, *China*;
2. *Key Laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi* 030006, *China*)

Abstract: Visual relationship detection is based on target recognition, and further detects the relationship between targets, which is a key technology of visual understanding and reasoning. However, the recall of few-shot tail relation detection is low due to the visual similarity of relation labels and the problem of data imbalance. In order to improve the detection effect of the tail relationship, this paper divides the relationship tags into coarse and fine-grained to construct a hierarchical representation of tags, and proposes a visual relationship detection model based on the tag hierarchy. The model uses the similarity between visual relationships and the bias of the data to build a hierarchical representation of relationship labels, so as to distinguish between coarse-grained relationships and fine-grained relationships, so that the tail relationships can be structured from coarse-grained to fine-grained and get more attention. At the same time, the loss function is designed according to the nature of the label hierarchy. The loss function learns the differences between different category relationships layer by layer through structured information, so that the model can better detect the fine-grained relationship in the tail. The effect of the proposed model in detecting tail relationships is verified in the public datasets Visual Relationship Detection (VRD) and Visual Genome (VG), respectively. Compared with the existing models, the average recall rates mR@20, mR@50 and mR@100 are improved by 0.62%, 1.57% and 2.47% in the VRD dataset, and 0.67%, 0.83% and 1.15% in the VG dataset, respectively.

Key words: visual relationship detection; tag hierarchical representation; long-tailed distributions; coarse-grained relationship; fine-grained relationship
Foundation Item(s): National Key Research and Development Program of China (No.2020AAA0106100); National

收稿日期:2022-09-14;修回日期:2023-04-20;责任编辑:李勇锋

Natural Science Foundation of China (No.62176145)

1 引言

视觉关系检测(visual relationship detection)[1~3]在图像理解中起着重要的作用,如图像描述[4,5]、视觉问答[6]等. 其具体任务是给定一副图像,检测并定位出图像中的对象,同时预测出对象之间的关系,通常采用主体-关系-客体三元组表示,比如“person-walk on-street”.

目前,视觉关系检测任务经典模型均采用对象-关系检测框架[7~9],对象检测模块检测并定位图像中的对象,关系检测模块提取对象的视觉特征、标签特征和位置特征,通过融合三种模态的特征来预测对象对之间的关系. 现有的方法虽然在视觉关系检测上取得了较好的效果,但由于关系样本长尾分布的影响,对少样本的关系检测结果并不理想[10]. 为了解决关系样本长尾分布带来的问题,一些研究者提出利用关系的先验知识来提高关系的少样本学习能力和关系预测精度. Lu等[1]首次提出了不同模态特征融合,通过预训练word2vec对检测出的对象标签进行编码,预测出关系的先验概率,并对视觉关系检测模型进行调整,以此提高尾部关系检测的准确率. Yu等[11]从外部常识库维基百科中检索丰富的文本知识,结合数据集内部的文本三元组知识,计算出确定对象对的视觉关系三元组的先验概率分布,以提高关系预测精度. Hwang等[12]首先通过对每一张图片中关系三元组统计来进行关系张量的构建,然后使用张量分解算法得到具有丰富信息的先验知识,用来提高关系预测的精度.

上述模型均聚焦于关系的先验概率来提高关系的检测效率,但是需要对关系文本数据进行额外的收集和处理,而且忽视了关系之间在视觉上的相似性导致了少量样本的尾部关系被错误的预测为样本丰富的头部关系. 如图1(*a*)所示,由于尾部关系“walk on”和头部关系“on”在视觉上具有相似性,以及头部关系具有丰富的样本而尾部关系只有少量的训练样本,模型错误预测图像中的关系为“person-on-street”. 此外,进一步分析模型的预测结果发现模型对“walk on”、“stand on”和“across”等尾部关系的区分效果也并不理想.

为了解决上述问题,本文首先将关系标签进行粗细粒度划分,其中,粗粒度关系是包含着丰富样本的头部关系,如:“on”、“next to”和“above”等,细粒度关系为与粗粒度关系同概念中的其他关系,细粒度关系通常样本稀少但具有更丰富信息,如相对于粗粒度关系“on”的细粒度关系有“stand on”和“walk on”等;然后根据粗粒度关系和细粒度关系在视觉上的相似性以及数据本身具有的偏见性,将粗粒度关系标签和细粒度关系标签形成层次结构表示,提出了基于标签层次结构的视觉关系检测模型. 如图1(*b*)所示关系标签的层次结构表示. 首先区分显著不同的关系,如“on”和“next to”;然后关注视觉上具有相似性的粗粒度和细粒度关系,如“on”和“walk on”;最后区分容易混淆的细粒度关系,如“walk on”和“stand on”. 为了更好的发挥模型对细粒度关系的检测效果,本文借鉴Yu等[13]方法构建了适合标签层次结构的损失函数,该损失函数使视觉关系检测模型不再一次性区分所有关系之间的细节差异,而是通过关系标签层次结构,逐层学习不同关系大类之间的特征以及粗粒度关系与细粒度关系之间的差异,以提高模型对细粒度关系的预测效果.

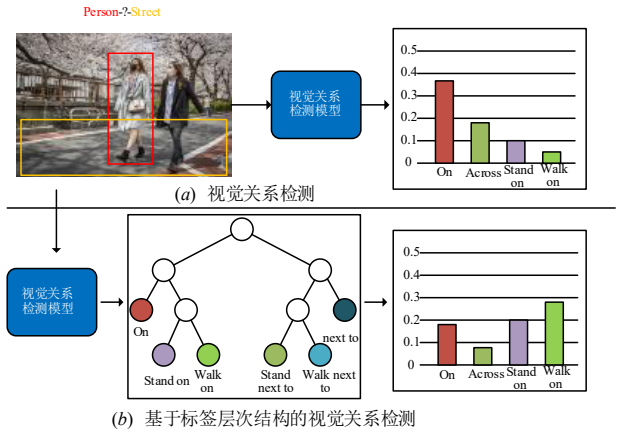

图1　关系检测示例分析

为了验证标签层次结构对解决细粒度关系检测的有效性,本文引入平均召回率[14](Mean Recall@K,mR@K)作为评价指标,该评价指标首先对每个关系类的召回率进行单独计算,然后计算得到每个类的召回率进行平均,即将所有的关系类视为同等重要. 在Visual Relationship Detection(VRD)[1]和Visual Genome(VG)[3]两个公开数据集上的实验结果表明本文方法的平均召回率均优于现有模型的平均召回率.

2 相关工作

为了提高计算机对图片的理解能力,视觉关系检测任务应运而生,许多现有的工作已经尝试使用视觉关系来促进特定的高级任务,如图像描述[4,5]、场景图生成[15,16]、视觉常识推理[17,18]和视觉问答[6]等.

早期的视觉关系检测任务只能预测少量的关系,Galleguillos等[18]利用对象对的共现性以及对象位置来预测对象对之间的四类空间关系:“above”、“below”、“inside”和“around”. Desai等[19]利用对象对的联合边界框的视觉特征和位置特征来预测8种常见的人类行为

关系. Sadeghi 等[2]首次引入视觉关系的三元组表示,利用一个关系三元组表示一个关系类别,将视觉关系检测转化为一个分类任务. 然而,当对象类别和关系类别增加时,视觉关系三元组的类别也非常的庞大.

一些视觉关系检测研究通过利用局部或全局上下文信息,即对象或者图像中的其他视觉关系作为上下文信息来协助预测关系. Zhuang 等[20]使用对象的词向量作为关系的语义上下文信息来协助预测对象间的关系,同时将上下文应用到视觉上的注意力机制中,让不同的对象能够关注不同的图像区域. Chen 等[21]将每张图像划分为若干个场景,以此作为全局上下文信息,并通过引入自注意力机制进行全局消息传递. Wang 等[10]利用一个包含特征级注意力机制的网络提取对象和关系的嵌入,由此产生的表示可以捕获更多潜在特征,再通过对偶图神经网络来传递和聚合关系和对象的上下文信息,以此提高关系识别的能力.

另外,一些视觉关系检测通过引入新的特征来提高视觉关系检测的效果. Zhou 等[7]提出在视觉关系检测的每个阶段深入挖掘和利用对象对的相对位置特征,提高候选对象对和关系检测的能力. Sharifzadeh 等[22]引入图像的深度图,通过深度图特征获得对象关系中有价值的信息. Zhang 等[23]将对象和关系的特征映射到统一低维空间中,使对象和关系表达为向量的转换. Han 等[24]利用注意力网络来观察图像中尺寸较小的对象,同时构建主体-客体层来区分对象对中的主体和客体,以此达到更精确的效果. Zhan 等[8]引入重要性检测作为目标检测和关系检测的补充,其中重要性检测是用于识别具有重要性的对象对. Yin 等[25]尝试构建内层次树(IH-tree)来解决歧义和嘈杂的对象和谓词注释. Mi 等[26]通过构建一种层次图注意网络来捕获对象级和三元组级的依赖关系,对象级层次图旨在捕获对象之间的交互,三元组级层次图用来构建关系三元组之间的依赖关系.

上述方法通过添加辅助信息和融合上下文信息一定程度上提高视觉关系检测效果,但在长尾问题上的表现依旧不理想;此外,在层次信息表示中也仅仅限于对象级的层次关系,忽略了关系之间的相关性. 因此,本文根据视觉关系检测模型的偏见性和关系之间的相关性形成关系标签的层次结构表示,以此区分粗粒度关系和细粒度关系. 同时,利用标签层次结构的损失函数,使模型更容易学到尾部关系的特征,从而来改进训练样本长尾分布带来的影响,得到更多有利于高级任务的细粒度关系.

3 基于标签层次结构的视觉关系检测方法

对于给定的图像 I,视觉关系检测任务是输出图像中对象对之间存在的关系,通常采用三元组来表示视觉关系 $<s, r, o>$. 具体概率模型如式(1)所示:

$$P(r)=P(r|s,o)P(s|B_s)P(o|B_o) \tag{1}$$

其中,s,o 和 r 分别表示主体,客体和关系. B_s 和 B_o 分别表示主体和客体的边界框. 主体和客体组成一个对象对. $P(s|B_s)$ 和 $P(o|B_o)$ 分别表示主体和客体的边界框属于某一个对象类别的概率. $P(r|s,o)$ 表示主客体组成的对象对之间存在的关系属于某一个关系类别的概率.

本文提出了基于标签层次结构的视觉关系检测模型. 首先,模型输入一张图片并经过预训练的目标检测模块,得到图像中的候选对象边界框以及候选对象的类别标签,具体采用 Faster-RCNN[27] 预训练模型作为目标检测模块. 然后进行视觉、位置、语言三种特征的融合,根据关系标签的层次结构以及标签层次结构的损失函数进行模型的训练. 具体模型分为多模态特征融合、关系标签层次结构构建、标签层次结构的损失函数三个部分,如图 2 所示.

3.1 多模态特征融合

本文采用 Faster-RCNN 预训练模型进行目标检测. 其中,用区域生成网络(Region Proposal Network,RPN)生成候选窗口,每张图片生成 300 个候选窗口. 候选区域池化(Regions Of Interest Pooling,ROI Pooling)[27]利用 RPN 生成的候选区域和 Faster-RCNN 网络最后一层得到固定大小的候选特征图. 采用全连接层进行目标识别和定位. 目标检测后模型分别从检测到的边界框和标签提取视觉特征、位置特征和语言特征. 最后,利用多模态特征融合策略来融合三种特征.

(1)视觉特征:该特征描述了对象的类别特征和对象在不同情况下的细节. 在关系预测模块,一般通过对特征图进行 ROI Pooling 操作提取视觉特征. 如图 2 所示,本文使图像经过以 Faster-RCNN 为骨干网络的卷积神经网络后,得到整张图像的特征图. 然后给定一对检测对象 s,o 的边界框 B_s 和 B_o,以及包含两个对象的最小边界框,使用 ROI Pooling 获得两个对象以及两对象之间关系的特征图. 将对象和关系的特征图经过卷积层和全连接层获得关系的视觉特征,再经过全连接层和对象的视觉特征融合形成最终的视觉特征. 视觉特征主要包括主体 s 和客体 o 视觉特征以及关系的视觉特征,表示为 $R_{\text{viso}}(s,o)$.

(2)位置特征:该特征主要指对象对的相对位置特征. 相对位置特征代表关系中两个对象之间的相对位置. 相对位置特征表示如式(2)所示.

$$R_{\text{loc}}(s,o)=\begin{bmatrix}\dfrac{x_s^1-x_u^1}{W_u},\dfrac{y_s^1-y_u^1}{H_u},\dfrac{x_s^2-x_u^1}{W_u},\dfrac{y_s^2-y_u^1}{H_u}\\ \dfrac{x_o^1-x_u^1}{W_u},\dfrac{y_o^1-y_u^1}{H_u},\dfrac{x_o^2-x_u^1}{H_u},\dfrac{y_o^2-y_u^1}{H_u}\end{bmatrix} \tag{2}$$

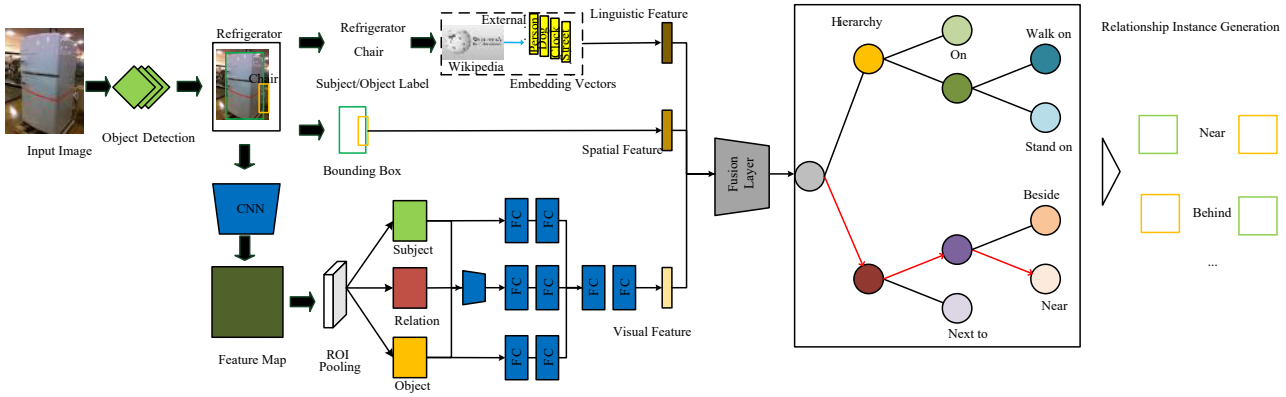

图 2　基于标签层次结构的视觉关系检测模型

其中,$[x_i^1, y_i^1, x_i^2, y_i^2]$, $i=s,o,u$ 分别表示主体和客体以及主客体的联合区域的边界框坐标. W_u 和 H_u 分别表示主客体的联合区域的边界框的宽和高.

(3)语言特征:该特征从语言知识中获得对象之间的相似性,而语言知识中获得对象之间的相似性很难从视觉外观和空间位置中获得. 在本文的视觉关系检测模型中,用词向量作为对象的语言特征. 首先,从目标检测模块获取对象类别. 然后,抽取外部语言特征,本文采用预训练的 word2vec 去获得对象对中主体 s 和客体 o 的语义表示. 语言特征表示为: $R_l(s,o)=[\text{word2vec}(s), \text{word2vec}(o)]$.

由于不同模态特征的维度差异很大,高维特征(如 4 096 维的视觉特征)很容易掩盖低维特征(如 8 维的位置特征). 为了缓解上述问题,将各模态的特征转化到相同的维度;随后再拼接为多模态特征. 特征融合表示如式(3)所示:

$$R_r = R_{\text{vis}} \cdot f(R_l, \theta_1) \cdot f(R_{\text{loc}}, \theta_2) \tag{3}$$

其中,R_r 为多模态特征融合后的表示,θ_1, θ_2 表示全连接层的学习权重.

3.2　关系标签层次结构的构建

现有的基于标签层次结构的方法侧重于概念上的相似性,而不是视觉上的相似性. 例如,根据 WordNet,由于鸟属于动物,所以鸟更接近猫而不是飞机. 然而,对于视觉相似性来说,情况正好相反. 由于鸟在天空中,所以视觉上更像飞机而不是猫. 因此,为了反映视觉相似性,本文提出了基于粗细粒度表示的标签层次结构构建方法.

具体在视觉关系检测中,不同关系类的样本会被预测为相同的关系,该类关系往往在视觉(如“stand next to”和“walk next to”)或高级语义(如“has”和“with”)上具有相似性. 本文将具有相同属性的尾部少样本关系聚集为一个大类,每一个大类都有一个抽象的概念,抽象概念用该大类中的粗粒度关系表示. 大类中的相同属性的尾部关系称为细粒度关系. 比如,尾部关系“walk on”和“stand on”在视觉上具有相似性,本文根据视觉关系检测模型的偏见性将“walk on”和“stand on”聚集到大类“on”中. 其中,“walk on”和“stand on”为细粒度关系,“on”为粗粒度关系. 具体构建过程分为以下三部分,如图 3 所示.

(1)获得关系样例概率分布. 对于所有的关系类 c_i 的样本,利用视觉关系检测模型本身具有的偏见性,计算出每个关系类 c_i 所有关系样本关系预测后的标签概率分布 $P_i=\left\{\frac{m_1}{n_i}, \frac{m_2}{n_i}, \frac{m_3}{n_i}, \cdots, \frac{m_k}{n_i}\right\}$. 其中,$n_i$ 表示关系类 c_i 的样本个数,m_k 表示关系类 c_i 的样本预测为第 k 个关系的个数. 由于关系样本的长尾分布以及部分尾部关系和头部视觉上的相似性,尾部关系样本的概率分布往往为头部关系. 如图 3(Step A),尾部关系“stand on”和“walk on”样本关系预测的概率分布中最高的为头部关系“on”.

(2)子层次结构构建. 在该过程将关系区分为粗粒度关系和细粒度关系. 对于关系类 c_i 的所有样本,关系预测的标签概率分布中最高的关系类 c_j 作为关系类 c_i 汇聚的大类的概念关系. 如果关系类 c_i 的所有样本关系预测的标签概率最高的关系类也为 c_i,则关系类 c_i 的大类的概念关系为其自身. 如图 3(Step A, Step B)所示,构建完所有的关系类后,关系“stand on”和“walk on”的所有样本关系预测的标签概率分布中概率最高的为“on”,所以“on”为“stand on”和“walk on”聚集的大类的概念关系. 子层次结构构建完毕后,大类的概念关系即为所有关系中样本丰富的粗粒度关系,其余关系为在视觉上与大类的概念关系具有相似性的少样本的细粒度关系.

(3)子层次结构聚合. 如图 3(Step C)所示,关系标签层次结构通过聚合最终变为四层,每层将关系归纳为比下一层中的关系更粗的关系. 具体而言,层次结构

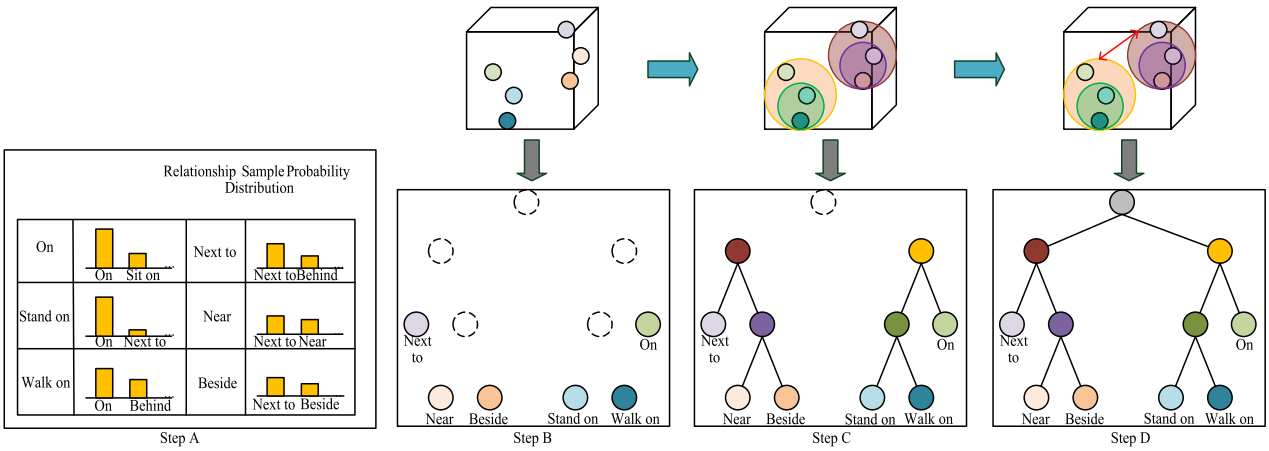

图 3 层次结构构建

的第一层包含一个虚拟节点. 虚拟节点不代表任何一个关系标签,只是一个模糊的概念. 层次结构的第二层用来区分输入的关系属于哪一个大类,该层包含了N个虚拟节点,代表着通过计算所有关系类c_i样本的标签概率分布产生的N个子层次结构. 层次结构的第三层主要用来区分输入的关系是用粗粒度的头部关系还是用细粒度的尾部关系进行描述. 第三层用于区分样本丰富的粗粒度关系“on”和样本稀少的细粒度关系“walk on”和“stand on”. 该层包含之前生成的N个子层次结构中的粗粒度的父类关系以及用于连接下一层的N个虚拟结点. 层次结构的第四层用于区分因为视觉相似性容易混淆的少样本的细粒度关系,如“walk on”和“stand on”,该层包含之前产生的子层次结构中的所有细粒度关系.

3.3 构建关系标签层次结构的损失

针对上述关系层次结构标签,提出了视觉关系检测模型损失函数,以此来使模型学习不同大类之间的特征,并检测细粒度的尾部关系. 该损失使模型可以根据标签层次结构从粗粒度关系中区分出细粒度关系. 具体分为两个互相补充的部分:基于标签路径的(label-path-based, LP)损失和类平衡(class-balanced, CB)损失[28]. 其中,基于标签路径的损失通过从粗粒度到细粒度来进行关系分类. 类平衡损失通过重新加权来消除模型的偏见.

关系标签层次结构的损失函数如式(4)所示:

$$L=\lambda \mathrm{LP}+\mathrm{CB}_{\mathrm{sigmoid}} \tag{4}$$

其中,λ是平衡权重,LP表示基于标签路径的损失,$\mathrm{CB}_{\mathrm{sigmoid}}$表示类平衡损失.

对于基于标签路径的损失,如图2层次结构中带箭头的线所示,标签层次结构中用正确路径L_{path}跟踪从虚拟根结点到关系结点的路径. 标签结构的正确路径表示对于一个训练样例在标签层次结构中正确的分类路径,对于给定带有一个正确路径L_{path}的样例,计算标签层次结构各层交叉熵损失以及正确样例的路径损失,得到基于标签路径的损失如式(5)所示:

$$\mathrm{LP}=-\sum_{i\in L_{\mathrm{path}}} w_i \log\left(\frac{\exp\left(Z_i\right)}{\sum\limits_{Z_j\in B(i)}\exp\left(Z_j\right)}\right)+\mathrm{CE}\left(D_{\mathrm{lh}},\ D_{\mathrm{label}}\right) \tag{5}$$

其中,$B(i)$表示结点i的兄弟结点,路径概率$D_{\mathrm{lh}}=\left\{P(r)\right\}_{r=1}^{R}$,正确标签$D_{\mathrm{label}}$,$\mathrm{CE}(\cdot)$表示交叉熵. 每个结点的概率如式(6)所示:

$$Z_i=\begin{cases} p_r, & \mathrm{leaf}(i)=T\cap \mathrm{class}(i)=r \\ \dfrac{1}{\left|L(i)\right|}\sum\limits_{j\in L(i)} Z_j, & \mathrm{leaf}(i)=F \end{cases} \tag{6}$$

其中,对于给定关系类r的样本,带有偏见性的视觉关系检测模型预测的概率为p_r. $L(i)$表示结点i的所有叶子结点,$\mathrm{leaf}(i)=T$表示结点i为叶子结点,$\mathrm{class}(i)=r$表示结点i为关系类r. $\mathrm{leaf}(i)=F$表示结点i不是叶子结点.

对于每一个节点i的孩子节点$j\in C(i)$的概率计算如式(7)所示:

$$P\left(j|i\right)=\mathrm{Soft\,max}\left(Z_i\right)\left[j\right] \tag{7}$$

一个叶子结点表示一个关系类r,从根节点到类r叶子结点的路径为L_{path},已知节点i属于路径L_{path},遍历下一个节点$C_r(i)\in L_{\mathrm{path}}\cap C(i)$,概率表示为$P_L\left(C_r(i)|i\right)$,最终,对每一个关系类$r$路径概率表示如式(8)所示:

$$P_L(r)=\prod_{i\in L_{\mathrm{path}}} P_L\left(C_r(i)|i\right) \tag{8}$$

基于标签路径的损失使网络越过各个大类间关系的噪声,分离叶结点代表的类,并分离层次结构中的每一内部节点,学习特定大类的嵌入,从而产生更细粒度的预测.

类平衡权重表示标签层次结构中每个结点的平衡权重. 本文采用权重因子来计算每个结点的类平衡权重,并统一定义每个结点的类平衡权重,如式(9)所示:

$$w_i=\begin{cases}\dfrac{(1-\beta)}{(1-\beta^{n_i})}, & \text{leaf}(i)=T\\ \dfrac{\sum\limits_{j\in L(i)}\omega_j}{|L(i)|\sum\limits_{k\in B(i)}\omega_k}, & \text{leaf}(i)=F\end{cases} \tag{9}$$

其中超参数$\beta\in[0,1)$.

类平衡损失用于解决在不平衡数据中进行训练的问题. 对于给定一个样例的正确关系r,基于带有偏见的预测概率P_{pre}计算类平衡sigmoid损失函数如式(10)所示:

$$\text{CB}_{\text{sigmoid}}=-w_i\sum_{p_j\in P_{\text{pre}}}\log\left(\frac{1}{1+\exp(-p_j)}\right) \tag{10}$$

其中,w_i是关系类i的权重因子.

4　实验与分析

本文模型在两个公开数据集Visual Relationship Detection[1]和Visual Genome[3]进行实验. Visual Relationship Detection(VRD)数据集由5 000张图片,100种对象和70种关系组成. 共包含37 993对关系三元组. 数据集被分为4 000张训练图片和1 000张测试图片. Visual Genome(VG)数据集是最大的视觉关系检测数据集之一. VG数据集有多个版本,在本文中,使用Xu等人[16]提供的VG数据集包含常见的150种对象类和50种常见的关系类. 本文按Zellers等人[29]的方法将VG数据集分为94 854张训练图片和13 223张测试图片,同时,从训练集中取样5 000张图作为验证集.

4.1　任务

本文使用三种用于评估视觉关系检测模型的任务,分别是谓词检测(predicate detection)、短语检测(phrase detection)和关系检测(relation detection).

(1)在谓词检测中,输入一张图片以及一组带有真实边界框和对象标签的对象对,模型预测出对象对之间的谓词关系.

(2)在短语检测中,输入一张图片,模型输出一个三元组〈对象1,关系,对象2〉,如果定位的整个三元组的边界框(对象对的联合边界框)与真实三元组的边界框的交并比(Intersection Over Union,IOU)大于0.5,则预测正确.

(3)在关系检测中,输入一张图片,模型输出一组三元组〈对象1,关系,对象2〉以及对象1和对象2的边界框,输出的边界框同对象1和对象2的真实边界框的IOU都大于0.5,则表示预测正确.

4.2　评价指标

平均召回率(Mean Recall@K,mR@K)[14]:为了反映模型在尾部关系的检测效果,本文引入平均召回率作为评价指标. 该性能指标独立计算每个关系类别的召回率,然后对所有关系类别的召回率进行平均. 因此,每个关系类别的贡献是相等的. 定义平均召回率如式(11)所示:

$$\text{Mean Recall@K}=\sum_{p\in\tau_p}\frac{\text{Recall@K}(p)}{|\tau_p|} \tag{11}$$

其中,τ_p表示所有关系的集合. Recall@K(p)表示计算关系p的召回率. 对所有模型用mR@K进行了评估,$K=20,50,100$.

4.3　实验参数设置

本文实验在Ubuntu 16.04.7,4块NVIDIA 1080Ti卡组成的服务器上运行,选用python语言及Pytorch1.5深度网络框架. 以Faster-RCNN为主干网络,输入图片大小为800×600,图像批次为16,首先通过49卷积层和1个全连接层获得图片的特征图,再通过Faster-RCNN的区域生成网络生成候选对象的边界框,最后对候选对象的边界框进行分类,得到每一个边界框的对象类别标签. 利用Relu函数和指数衰减学习率的Adam优化器来训练视觉关系检测网络. 对于VRD数据集,初始学习率设置为0.000 3,每4 000步衰减0.5. 对于VG数据集,初始学习率设置为0.000 3,每35 000步衰减0.7. 经实验验证,λ设置为1,β设置为0.999.

4.4　实验设计

本文在VRD数据集上与2种视觉关系检测模型进行比较,分别是MF-URLN[8]和RLM[7];在VG数据集上与5种视觉关系检测模型进行比较,分别是VTransE[23]、MF-URLN[8]、RLM[7]、VRD-DM[22]和RelTransformer[21].

MF-URLN[8]:该模型提出了不确定关系对视觉关系检测存在一定的影响,构建了基于多模态特征的不确定关系学习网络. 在视觉关系检测方面取得了很大的改进.

RLM[7]:该模型通过深入挖掘和利用对象对视觉关系检测过程的每个阶段的相对位置,提取每个对象对的相对位置信息并将其编码为辅助特征,并引入门控图神经网络,利用相对位置挖掘关系的相关性,以提高对象对的提出和对相似关系区分的能力.

VTransE[23]：该模型将知识图谱中的关系表示方法TransE应用于视觉关系表示，首先将对象和关系的特征映射到统一低维空间中，使对象和关系表达为向量的转换，然后以此构建并训练模型，提高视觉关系检测的效果.

VRD-DM[22]：该模型提出利用深度图提供关于对象关系有价值的信息. 融合深度图特征、RGB图像特征、标签特征以及位置特征进行视觉关系检测.

RelTransformer[21]：该模型将每一张图像表示为一个完全连接的场景图，将场景重构为关系三元组和全局上下文，并通过自注意力机制将来自关系三元组和全局上下文中的信息传递给目标关系.

4.5 实验结果与分析

为了验证标签层次结构的有效性，本文进行了三部分实验，分别是与主流视觉关系检测方法的比较试验、消融实验和性能分析. 在消融实验中，首先，为了证明基于标签层次结构的损失和类平衡sigmoid损失组合具有互补优势并使模型达到最佳效果，用不同的损失函数训练视觉关系检测模型并在VRD数据集上进行实验；其次，为了证明本文提出的标签层次结构的构建方法的有效性，将不同形式的层次结构在VRD数据及上进行了实验；最后，对参数λ和β的取值在VRD数据及上进行了实验. 在性能分析中，展示VRDMLH与现有模型RLM在尾部粗粒度关系的检测效果.

4.5.1 与主流视觉关系检测方法的比较

如表1所示，在VRD数据集中，本文提出的基于标签层次结构的视觉关系检测方法在谓词检测、短语检测和关系检测三方面的平均召回率都优于现有的模型的平均召回率. 在谓词检测中，与最先进模型RLM相比，VRDMLH的mR@20、mR@50和mR@100分别提高了1.69%、4.29%、5.35%. 说明在给定对象对的情况下，VRDMLH在尾部细粒度关系的检测效果要优于已有的模型并且可以检测出更多种关系. 在关系检测任务中，与最先进模型RLM相比，VRDMLH的平均召回率要优于RLM.

表1 VRD数据集上结果

方法	Predicate Detection	Phrase Detection	Relation Detection
	mR@20/50/100	mR@20/50/100	mR@20/50/100
MF-URLN	9.97/10.47/10.77	4.18/5.52/6.54	3.12/4.08/4.70
RLM	10.36/10.39/10.39	4.08/5.78/7.03	3.16/4.42/5.17
VRDMLH	**12.05/14.68/15.74**	**8.60/10.31/10.72**	**3.78/5.99/7.64**

如表2所示，在VG数据集中，本文提出的基于标签层次结构的视觉关系检测方法在谓词检测、短语检测和关系检测三方面的平均召回率都优于现有的模型. 在谓词检测中，与最先进模型VRD-DM相比，VRDMLH的mR@20、mR@50和mR@100分别提高了1.82%、2.32%、2.51%. 说明在给定对象对的情况下，VRDMLH在尾部细粒度关系的检测效果要优于已有的模型并且可以检测出更多种关系.

表2 VG数据集上结果

方法	Predicate Detection	Phrase Detection	Relation Detection
	mR@20/50/100	mR@20/50/100	mR@20/50/100
VTransE	11.58/14.73/15.82	6.65/8.23/8.72	3.67/4.96/6.00
MF-URLN	8.14/10.06/10.80	5.29/6.39/6.75	2.59/3.78/4.68
RLM	13.00/16.48/17.80	6.68/8.34/8.87	4.77/6.52/7.64
VRD-DM	16.40/20.74/22.72	—	—
RelTransformer	18.51/19.58/20.19	—	—
VRDMLH	**18.22/23.06/25.23**	**8.34/11.25/12.73**	**5.44/7.35/8.79**

4.5.2 消融实验

为了验证基于标签层次结构的类平衡损失与类平衡sigmoid损失结合具有互补优势，以及二者组合可以得到最优的效果，本文对视觉关系检测模型采用不同损失函数在VRD数据集上的关系检测效果进行了研究，如表3所示，LP表示仅使用基于标签层次结构的类平衡损失，$CB_{sigmoid}$表示仅使用类平衡sigmoid损失，$CB_{softmax}$表示仅使用类平衡softmax损失，TCB表示Yu等人[13]的基于CogTree的类平衡损失，Full Loss表示同时使用基于标签路径的损失和类平衡sigmoid损失. 通过实验说明基于标签路径的损失与类平衡损失结合具有互补优势，可以使模型训练达到最好的效果.

表3 VRD数据集上不同损失模型检测结果

损失函数	Predicate Detection	Phrase Detection	Relation Detection
	mR@20/50/100	mR@20/50/100	mR@20/50/100
LP	11.78/12.87/13.01	8.30/9.03/9.21	3.36/5.29/7.52
$CB_{sigmoid}$	9.75/12.62/13.08	7.06/8.57/9.11	3.50/5.50/7.02
$CB_{softmax}$	11.52/12.61/12.28	8.00/8.72/8.83	3.24/5.13/7.37
TCB	9.11/10.44/10.68	6.48/7.31/7.44	2.55/3.74/5.53
Full Loss	**12.05/14.68/15.74**	**8.60/10.31/10.72**	**3.78/5.99/7.64**

为了进一步证实是关系标签层次结构一定程度上解决了因数据的长尾分布导致的关系检测模型只能检测到头部粗粒度关系的问题，本文将带有标签层次结

构的视觉关系检测模型(VRDMLH)和没有标签层次结构的视觉关系检测模型(Visual Relationship Detection Model based on Label Hierarchy, VRDMNLH)在VRD数据集上进行了实验,如表4所示,与VRDMNLH相比,在谓词检测上,VRDMLH的mR@20、mR@50和mR@100分别提高了3.08%、4.58%、5.44%. 在短语检测上,VRDMLH的mR@20、mR@50和mR@100分别提高了2.00%、2.98%、3.16%. 在关系检测上,VRDMLH的mR@20、mR@50和mR@100分别提高了0.70%、2.08%、2.46%. 通过实验说明标签关系层次结构可以有效地使模型区分粗粒度关系和细粒度关系,解决模型对细粒度关系样本预测错误的问题.

表4　VRD数据集上标签层次结构不同构建方式的结果

模型	Predicate Detection	Phrase Detection	Relation Detection
	mR@20/50/100	mR@20/50/100	mR@20/50/100
Fuse Subtree	7.95/11.89/13.20	6.02/8.32/9.16	2.61/4.61/6.41
Fuse Layer	7.27/10.32/11.11	6.14/8.81/9.70	2.78/4.72/6.87
CogTree	10.48/11.59/11.75	7.48/8.67/9.44	3.50/5.85/7.02
VRDMNLH	8.97/10.10/10.30	6.60/7.33/7.56	3.08/3.91/5.18
VRDMLH	**12.05/14.68/15.74**	**8.60/10.31/10.72**	**3.78/5.99/7.64**

同时,为了验证本文构建标签层次结构中构建关系大类和区分粗粒度关系和细粒度关系的有效性,本文提供了融合所有子树的层次结构的构建方法,即标签层次结构的第二层为所有关系;融合了粗粒度关系和细粒度关系的构建方法,即将标签层次结构的第三四层融合. 以及Yu等[13]提出的CogTree的构建方式. 如表4所示,Fuse Subtree表示融合所有子树的层次结构构建方法,Fuse Layer表示融合了粗粒度关系和细粒度关系的构建方法. 本文提出的利用视觉关系检测模型的偏见性构建的标签层次结构主要将属于同一大类关系组织在一个子层次结构中,并将一个子层次结构的关系从粗粒度到细粒度分属于不同层次中. 从表4中可以得到当违反了构建层次结构的规定时,模型的检测效果会发生显著的下降.

此外,本文对式(4)和式(7)中的参数λ和β进行了评估,以证实$\lambda=1$,$\beta=0.999$时,VRDMLH能达到最佳性能. 对参数$\lambda=\{0.4,0.7,1,1.3,1.6\}$,$\beta=\{0,0.9,0.99,0.999\}$在VRD数据集上进行了谓词检测实验. 如图4所示,当β取定值时,$\lambda=1$时模型效果最佳;当λ取定值时,$\beta=0.999$时模型效果最佳,图中明显得出模型在$\lambda=1$,$\beta=0.999$时,模型各项指标达到最佳性能.

4.5.3　效果分析

为了展示基于标签层次结构的视觉关系检测模型对尾部细粒度关系的检测效果,图5中统计了在VRD数据集中RLM和VRDMLH分别在尾部细粒度关系样本上的检测效果,其中蓝色表示RLM模型,橘黄色表示

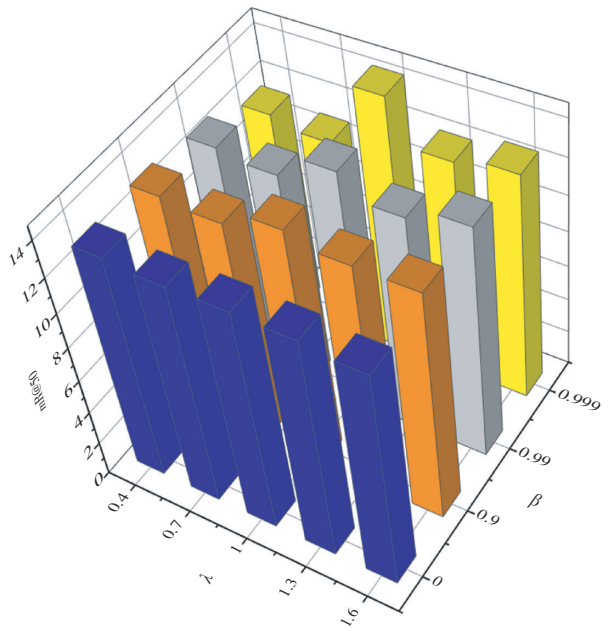

图4　不同λ和β效果对比

VRDMLH模型. 可以明显的看出,RLM不具备部分尾部细粒度关系的检测能力,比如,在“sleep on”、“lying on”和“drive”的样本上的检测效果为0,而VRDMLH可以在尾部细粒度关系的部分样本上检测成功. 同时,在一些RLM可以检测到的尾部关系,VRDMLH都有一定的提高,比如“hold”、“ride”和“fly”等. 但是,在图中一些头部粗粒度关系的部分样本中,VRDMLH的检测效果会有一定下降,主要是因为VRDMLH是通过标签的层次结构进行模型的训练,当某个大类的细粒度关系较多时,由于层次结构中虚拟结点概率的计算方式会导致模型更侧重于检测细粒度关系而非粗粒度关系. 对粗粒度关系样本检测为细粒度关系仅仅是检测结果与正确结果不符的错误,并不表示视觉中的关系是完

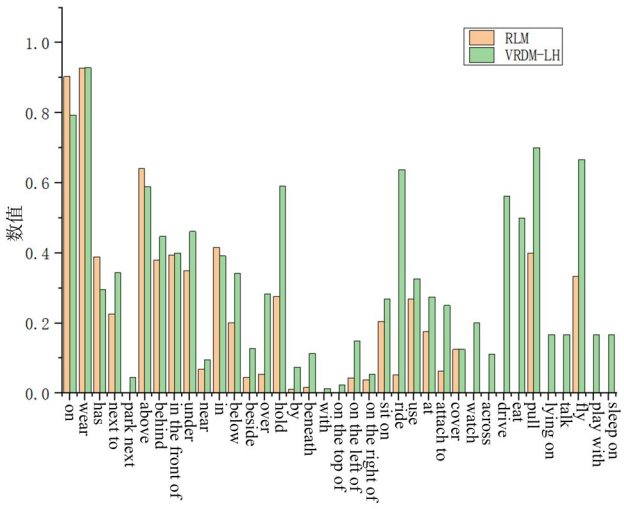

图5　RLM和VRDMLH在VRD上部分关系样本预测结果

全错误的. 而RLM是通过相对位置评估模块预测更多的位置头部关系,以此提高了头部关系检测的效果,虽然提高了检测效果,但对尾部关系的检测依旧不理想. 整体而言,RLM可以检测到更多的头部粗粒度样本的关系,而VRDMLH可以更均衡的检测到图片中的各类关系.

为了验证本文模型在构建标签层次结构时的合理性,用样例逐层分类的散点图展示模型的推理过程. 如图6所示为三层关系标签结构的生成过程(输入一个批次16张图片,生成105对关系样例),横轴表示关系样本中的关系序号,共70种关系,纵轴表示根据数据生成的层次结构中每一层的结点数. 具体为:第一层模型生成有10个虚拟节点,将所有对象对中的关系分在10个虚拟节点中;第二层对第一层中的节点进行分类以区分了粗粒度关系和细粒度关系,其中细粒度关系为新的虚拟节点;第三层中对第二层的虚拟节点再进行细粒度之间的划分. 整体结构通过基于标签路径的损失不断优化,使模型具有区分粗、细粒度关系样本的能力. 为了更清楚的显示推理过程,抽取了每一层中的样例进行分析,n00000001表示第一层中10节点对应的样例分布图,n10000004表示在第二层中所对应的第一层第4个结点所划分的虚拟结点,n20000004表示在第三层中所对应的第二层第4个节点所对应的划分. 从图中可以看出,105对关系样例在第二层中有一些关系生成为粗粒度关系,在第三层中被划分成更细粒度的关系,动态展示了本文方法粗细粒度划分的推理过程.

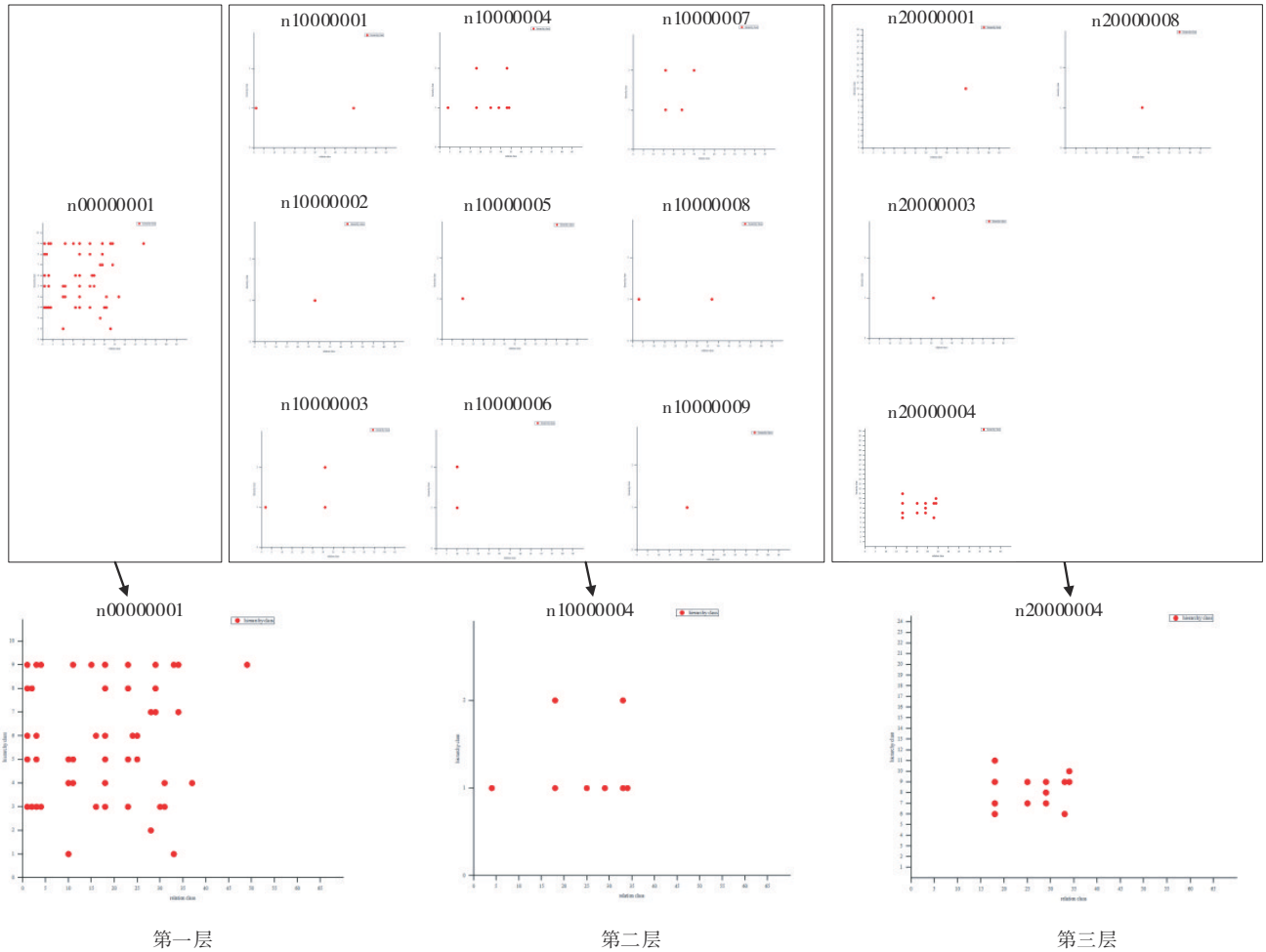

图6 层次结构训练样例分布图

5 结论

本文在视觉关系检测中提出将关系标签形成层次结构表示,同时,提出基于标签层次结构的损失,该损失通过长尾分布的数据来预测不带有偏见的关系. 利用带有偏见预测视觉关系检测模型,将独立的关系以层次结构的形式进行组织,关系标签的层次结构由粗粒度到细粒度包含了多层对应的关系,提出基于标签层次结构的视觉关系检测模型. 实验结果表明本文方法可以有效地缓减粗粒度关系错误地预测为细粒度关系的问题. 接下来会探索如何利用图文常识对标签的

层次结构进行优化.

参考文献

[1] LU C W, KRISHNA R, BERNSTEIN M, et al. Visual relationship detection with language Priors[C]//Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 852-869.

[2] SADEGHI M A, FARHADI A. Recognition using visual phrases[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2011: 1745-1752.

[3] 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.

[4] 周东明, 张灿龙, 李志欣, 等. 基于多层级视觉融合的图像描述模型[J]. 电子学报, 2021, 49(7): 1286-1290.
ZHOU D M, ZHANG C L, LI Z X, et al. Image captioning model based on multi-level visual fusion[J]. Acta Electronica Sinica, 2021, 49(7): 1286-1290. (in Chinese)

[5] 罗会兰, 郭敏杰, 孔繁胜. 一种基于多级空间视觉词典集体的图像分类方法[J]. 电子学报, 2015, 43(4): 684-693.
LUO H L, GUO M J, KONG F S. An image classification method based on multiple level spatial visual dictionary ensemble[J]. Acta Electronica Sinica, 2015, 43(4): 684-693. (in Chinese)

[6] 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.
YU J, WANG L, YU Z. Research on visual question answering techniques[J]. Journal of Computer Research and Development, 2018, 55(9): 1946-1958. (in Chinese)

[7] ZHOU H, ZHANG C Y, HU C P. Visual relationship detection with relative location mining[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 30-38.

[8] ZHAN Y, YU J, YU T, et al. On exploring undetermined relationships for visual relationship detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 5128-5137.

[9] JUNG J, PARK J. Visual relationship detection with language prior and softmax[C]//2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS). Piscataway: IEEE, 2018: 143-148.

[10] WANG W T, WANG M, WANG S, et al. One-shot learning for long-tail visual relation detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12225-12232.

[11] YU R C, LI A, MORARIU V I, et al. Visual relationship detection with internal and external linguistic knowledge distillation[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 1068-1076.

[12] HWANG S J, KIM H J, RAVI S N, et al. Tensorize, factorize and regularize: Robust visual relationship learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1014-1023.

[13] YU J, CHAI Y, WANG Y J, et al. CogTree: Cognition tree loss for unbiased scene graph generation[C]// International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 1274-1280.

[14] TANG K, ZHANG H, WU B, et al. Learning to compose dynamic tree structures for visual contexts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 6612-6621.

[15] 吴绿, 张馨月, 唐茉, 等. Focus+Context语义表征的场景图像分割[J]. 电子学报, 2021, 49(3): 596-604.
WU L, ZHANG X Y, TANG M, et al. Focus+Context semantic representation in scene segmentation[J]. Acta Electronica Sinica, 2021, 49(3): 596-604. (in Chinese)

[16] XU D, ZHU Y, CHOY C B, et al. Scene graph generation by iterative message passing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5410-5419.

[17] CUI W, LAN Y, PANG L, et al. Beyond language: Learning commonsense from images for reasoning[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 4379-4389.

[18] GALLEGUILLOS C, RABINOVICH A, BELONGIE S. Object categorization using co-occurrence, location and appearance[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2008: 1-8.

[19] DESAI C, RAMANAN D. Detecting actions, poses, and objects with relational phraselets[C]//European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2012: 158-172.

[20] ZHUANG B H, LIU L Q, SHEN C H, et al. Towards context-aware interaction recognition for visual relationship detection [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 589-598.

[21] CHEN J, AGARWAL A, ABDELKARIM S, et al. RelTransformer: A transformer-based long-tail visual relationship recognition[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 19507-19517.
[22] SHARIFZADEH S, BAHARLOU S M, BERRENDORF M, et al. Improving visual relation detection using depth maps [C]//2020 25th International Conference on Pattern Recognition. Piscataway: IEEE, 2021: 3597-3604.
[23] ZHANG H, KYAW Z, CHANG S F, et al. Visual translation embedding network for visual relation detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5532-5540.
[24] HAN C, SHEN F, LIU L, et al. Visual spatial attention network for relationship detection[C]//Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM, 2018: 510-518.
[25] YIN G J, SHENG L, LIU B, et al. Zoom-Net: Mining deep feature interactions for visual relationship recognition[C]// Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 330-347.
[26] MI L, CHEN Z. Hierarchical graph attention network for visual relationship detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 13886-13895.
[27] REN S Q, HE K M, GIRSHICK R, 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.
[28] CUI Y, JIA M, LIN T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 9268-9277.
[29] ZELLERS R, YATSKAR M, THOMSON S, et al. Neural motifs: Scene graph parsing with global context[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5831-5840.

作者简介

王元龙 男，1983 年 5 月出生于山西省大同市．现为山西大学计算机与信息技术学院副教授．
E-mail: ylwang@sxu. edu. cn

雷 鸣 男，1999 年 2 月出生于山西省运城市．现为山西大学计算机与信息技术学院研究生．