软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(9):4207−4222 [doi: 10.13328/j.cnki.jos.007303] [CSTR: 32375.14.jos.007303]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

提升隐式场景下短语视觉定位的因果建模方法*

赵嘉宁, 王晶晶, 罗佳敏, 周国栋

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 王晶晶, E-mail: djingwang@suda.edu.cn

摘　要: 短语视觉定位是多模态研究中一个基础且重要的研究任务, 旨在预测细粒度的文本短语与图片区域的对齐关系. 尽管已有的短语视觉定位方法已经取得了不错的进展, 但都忽略了文本中的短语与其对应图片区域的隐式对齐关系 (即隐式短语-区域对齐关系), 而预测这种关系可以有效评估模型理解深层多模态语义的能力. 因此, 为了有效建模隐式短语-区域对齐关系, 提出一种隐式增强的因果建模短语视觉定位方法. 该方法使用因果推理中的干预策略来缓解浅层语义所带来的混淆信息. 为评估模型理解深层多模态语义的能力, 标注一个高质量的隐式数据集, 并进行大量实验. 多组对比实验结果表明, 所提方法能够有效建模隐式短语-区域对齐关系. 此外, 在这个隐式数据集上, 所提方法的性能优于一些先进的多模态大语言模型, 这将进一步促进多模态大模型更多的面向隐式场景的研究.

关键词: 隐式短语-区域对齐关系; 因果推理; 短语视觉定位

中图法分类号: TP18

中文引用格式: 赵嘉宁, 王晶晶, 罗佳敏, 周国栋. 提升隐式场景下短语视觉定位的因果建模方法. 软件学报, 2025, 36(9): 4207–4222. http://www.jos.org.cn/1000-9825/7303.htm

英文引用格式: Zhao JN, Wang JJ, Luo JM, Zhou GD. Implicit-enhanced Causal Modeling Method for Phrasal Visual Grounding. Ruan Jian Xue Bao/Journal of Software, 2025, 36(9): 4207–4222 (in Chinese). http://www.jos.org.cn/1000-9825/7303.htm

Implicit-enhanced Causal Modeling Method for Phrasal Visual Grounding

ZHAO Jia-Ning, WANG Jing-Jing, LUO Jia-Min, ZHOU Guo-Dong

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Phrasal visual grounding, a fundamental and critical research task in the field of multimodal studies, aims at predicting fine-grained alignment relationships between textual phrases and image regions. Despite the remarkable progress achieved by existing phrasal visual grounding approaches, they all ignore the implicit alignment relationships between textual phrases and their corresponding image regions, commonly referred to as implicit phrase-region alignment. Predicting such relationships can effectively evaluate the ability of models to understand deep multimodal semantics. Therefore, to effectively model implicit phrase-region alignment relationships, this study proposes an implicit-enhanced causal modeling (ICM) approach for phrasal visual grounding, which employs the intervention strategies of causal reasoning to mitigate the confusion caused by shallow semantics. To evaluate models' ability to understand deep multimodal semantics, this study annotates a high-quality implicit dataset and conducts a large number of experiments. Multiple sets of comparative experimental results demonstrate the effectiveness of the proposed ICM approach in modeling implicit phrase-region alignment relationships. Furthermore, the proposed ICM approach outperforms some advanced multimodal large language models (MLLMs) on the implicit dataset, further promoting the research of MLLMs towards more implicit scenarios.

Key words: implicit phrase-region alignment; causal inference; phrasal visual grounding

在连接人类和机器智能方面, 视觉场景和自然语言描述的跨模态理解发挥着至关重要的作用[1], 其中一个主要的问题是如何建立视觉区域与相关短语描述间的细粒度对齐关系, 这通常被称为短语视觉定位 (phrasal visual

* 基金项目: 国家自然科学基金 (62006166, 62076175, 62076176); 江苏高校优势学科建设工程
收稿时间: 2023-11-01; 修改时间: 2024-07-09; 采用时间: 2024-10-09; jos 在线出版时间: 2025-02-26
CNKI 网络首发时间: 2025-02-27

© 中国科学院软件研究所 http://www.jos.org.cn

grounding, PVG) 任务[2]. 图 1(a)、(b) 给出了 PVG 任务的示例, 文本中每种颜色的短语在图像中均有相同颜色的区域框对应. 例如, (a) 中绿色的短语“一名小孩 (a small child)”对应绿色的区域框; (b) 中紫色的短语“浅棕色衬衫 (a light brown shirt)”对应紫色的区域框. 这种对应关系是很多视觉语言多模态任务的基础, 如图像描述 (image captioning)[3], 图像检索 (image retrieval)[4], 视觉问答 (visual question answering)[5]等.

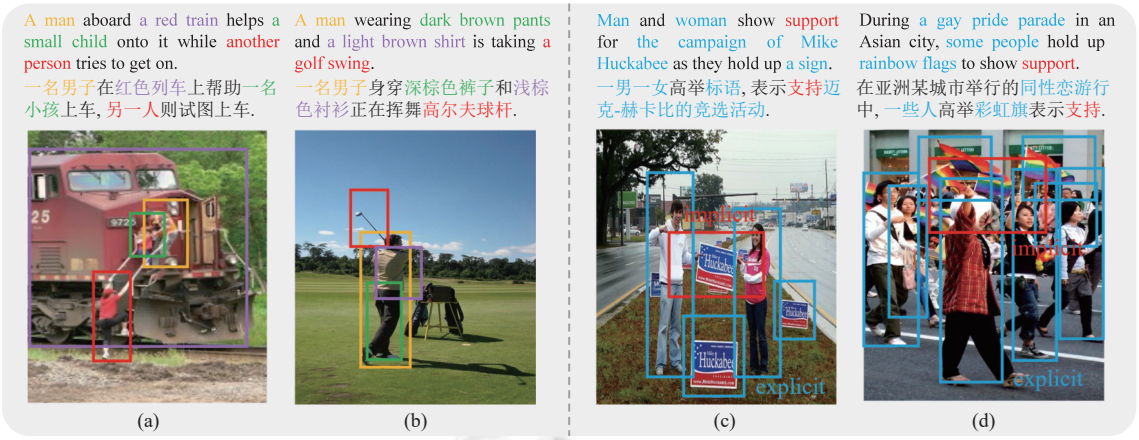


图 1 PVG 任务示例以及包含隐式关系 (implicit) 和显式关系 (explicit) 的图像-文本对

近年来, PVG 任务的研究发展迅速, 其模型大多采用双分支网络分别提取图像特征和文本特征, 经过多模态特征融合后, 预测短语对应的边界框. 随着对比学习在多模态领域的应用[6], PVG 任务的性能也得以迅速提升[7,8]. 然而, 已有的 PVG 工作都只关注了短语和区域间有着浅层对应关系的样本 (即显式短语-区域对齐关系), 忽视了其他一些短语和区域间有着深层对应关系的样本 (即隐式短语-区域对齐关系). 例如, 在图 1(c)、(d) 中, 蓝色的短语“一男一女 (man, woman)”“标语 (a sign)”“同性恋游行 (a gay pride parade)”和“彩虹旗 (rainbow flags)”等在图像中有着蓝色区域框与其相对应, 模型很容易学习到这种浅层对应关系; 而对于红色的“支持 (support)”短语, 它们与图像中的区域具有深层的对应关系, 需要模型进一步理解深层语义, 学习到“举着手”的动作是一种表示“支持”的常识, 才能将“支持 (support)”与图 1(c)、(d) 中红色的区域框对应, 这对于已有的模型而言是异常困难的. 本文中, 我们将类似“支持 (support)”这种短语和区域间有着深层对应关系的短语-区域对定义为一种“隐式短语-区域对齐关系”, 简称为“隐式关系”; 反之, 若短语和区域间有着浅层对应关系, 我们将这类短语-区域对定义为“显式短语-区域对齐关系”, 简称为“显式关系”. 基于对现有数据集的观察和分析, 本文通过预标注总结了 4 种隐式关系: 1) 常识性理解; 2) 上下文理解; 3) 空间关系理解; 4) 数值信息理解, 并构建了一个面向隐式场景的 PVG 数据集用以评估模型深层多模态语义的理解能力 (详情请见第 4.1 节).

本文认为建模隐式关系是一个巨大的挑战, 而已有的工作大多将重心放在如何学习短语和区域之间的关联性, 在预测隐式短语的对应区域时, 往往不会考虑理解隐式短语的深层语义. 因此, 预测结果常会被其他一些浅层语义所混淆. 例如, “支持 (support)”所对应的区域为图 1(c)、(d) 中特殊的红色区域框 (即“人们举着手”的动作), 而“举着手”是人所特有的一种动作, 因此这种文本短语与图像区域的隐式对齐关系很容易被浅层语义 (如短语“一男一女 (man and woman)”和“一些人 (some people)”及其所对应的区域) 混淆, 导致预测结果出现偏差.

为了缓解浅层语义所带来的混淆问题, 本文受因果推理思想的启发[9], 设计了一种新的 PVG 方法来建模隐式关系. 该方法采用因果推理中的干预策略, 缓解了浅层语义会误导模型错误对齐隐式短语与其对应区域的问题. 具体而言, 本文提出了一种隐式增强的因果建模短语视觉定位方法 (implicit-enhanced causal modeling approach for phrasal visual grounding, ICM). 该 ICM 方法主要包含 3 个部分: 编码模块 (encoding block), 隐式感知的因果注意力模块 (implicit-aware causal attention module, ICA) 和隐式感知的优化模块 (implicit-aware optimization). 首先, ICM

© 中国科学院软件研究所 http://www.jos.org.cn

使用编码模块分别编码图像和文本特征; 然后, ICM 在 ICA 模块中采用因果推理中的前门调整策略对图像和文本特征进行融合与去混杂, 从而缓解了模型在预测隐式关系时会被浅层语义所混淆的问题; 最后, ICM 在隐式感知的优化模块中, 利用去混杂后的图像和文本特征预测文本中短语在图像中所处的区域. 相较于传统的 PVG 方法, ICM 更关注文本中语义较深且复杂的短语, 从而能够有效提升模型理解深层多模态语义的能力.

综上所述, 本文的贡献如下.

(1) 本文考虑了短语视觉定位 (PVG) 任务中的隐式短语-区域对齐关系问题, 并基于对现有数据集的分析构建了一个面向隐式场景的高质量隐式数据集, 用于帮助评估模型深层多模态语义的理解能力.

(2) 本文为短语视觉定位任务提出了一种新颖的隐式增强的因果建模短语视觉定位方法 ICM. 具体而言, 本文首先构建了因果图对 PVG 的显式和隐式短语-区域对齐关系进行了分析, 然后采用前门调整策略缓解模型在预测隐式关系时会被浅层语义混淆的问题.

(3) 本文发现 ICM 在我们构建的隐式数据集上的性能优于一些先进的多模态大语言模型, 这将进一步促进多模态大模型更多的面向隐式场景的研究.

1 相关工作

1.1 视觉定位

视觉定位 (visual grounding) 任务是多模态领域的常见任务之一, 按照是否要对语言描述中所有的短语进行定位, 可以进一步地将其划分为两个任务: 短语视觉定位 (PVG)[2]和目标指代理解 (REC)[10]. 对于给定的图像-文本对, 数据集中对于所有的短语都有标注. PVG 任务需要在图像中定位文本中提到的所有短语, 而 REC 任务只对数据集标注的一个短语进行定位. 本文主要关注 PVG 任务, 目前该任务的研究方法大致可以分为两种形式: 两阶段方法和单阶段方法[11].

两阶段方法将 PVG 任务的过程分为两个步骤: 1) 首先利用一个预训练的目标检测模型 (如 Faster R-CNN[12]等) 从图像中提取一组候选区域, 2) 然后将待对齐的短语与候选区域进行相似度排序, 返回相似度最高的区域. 例如, 早期的 MattNet[13]将待定位短语和图像分解为与主题, 位置和关系相关的 3 个模块化组件, 以建模细粒度相似度. Zhuang 等人[14]使用注意力机制重构了一个并行注意力网络, 来发现图像中被不同长度的语言描述所提及的区域. Yu 等人[15]发现现有的两阶段方法更注重多模态表示的生成和如何更好地对目标检测模型生成的候选区域排序. 基于此, 他们提出了多样化和鉴别性网络 DDPN 来改进候选区域的生成, 同时考虑了多样性和区分度. 还有一些工作利用图学习来更好地进行多模态对齐. 例如, Yang 等人[16]和 Wang 等人[17]提出了图注意力网络来完成 PVG 任务, 而 Yang 等人[18]则利用门控图卷积网络融合多模态信息, 提出了跨模态关系推理网络 CMRIN. 然而, 这些方法很大程度上依赖于预训练的目标检测模型的性能. 如果在第 1 阶段中没有生成与待定位短语对齐的候选区域, 第 2 阶段的排序和选择过程也无法输出正确的定位结果[19]. 并且, 在第 1 阶段中使用目标检测模型生成候选区域往往需要耗费大量的时间[20].

为了解决两阶段方法对预训练的目标检测模型极大的依赖性和生成候选区域需耗费大量时间的问题, 研究者们提出了单阶段方法, 其无需事先生成候选区域, 而是将图像特征和文本特征紧密融合为多模态特征, 并利用多模态特征图, 以滑动窗口的方式直接进行边界框的预测. FAOA[11]将文本编码为向量, 并将其与 YOLOv3[21]作为目标检测器提取的图像特征融合, 同时使用空间特征来增强视觉特征以完成 PVG 任务. RCCF[22]将 PVG 任务定义为关联过滤过程, 并选择相关热力图的峰值作为目标区域的中心. Yang 等人[23]为了解决 FAOA 在面对复杂的短语查询时定位性能不高的局限性, 设计了一个递归子查询构造网络 ReSC. LBYL-Net[24]设计了一个 landmark 卷积模块, 在语言描述的指导下传输视觉特征, 并对目标区域与其对应的上下文之间的空间关系进行编码. Liao 等人[25]提出了一种语言引导的视觉特征学习机制, 其中语言信息在一开始就用来指导视觉特征的提取, 从而充分利用了两种模态的信息. 随着 Transformer[26]的广泛应用, Deng 等人[27]最早提出了基于 Transformer 的端到端的 PVG 方法 TransVG. TransVG 是一个由多个 Transformer 堆叠的网络, 包括文本编码器 BERT[28], 视觉编码器 DETR[29]和多模

© 中国科学院软件研究所 http://www.jos.org.cn

态特征融合 Transformer. 还有一些最新的基于 Transformer 的单阶段方法专注于对视觉编码器分支的改进, 并结合多模态特征调整视觉特征. VLTVG[19]使用视觉-语言验证模块调整视觉特征, 并使用语言引导的上下文编码器聚合视觉上下文. QRNet[20]通过查询感知动态注意力 (QD-ATT) 机制和查询感知多尺度融合来调整视觉特征. 随着对比学习在多模态领域的应用, 以及多模态预训练的兴起, Kamath 等人[7]将 PVG 任务建模为一个调制检测任务, 提出了一种源自 DETR 检测器的新型框架 MDETR, 并采用对比学习的思想设计了一个新的损失函数有效学习短语和区域的对应关系. Li 等人[8]将目标检测任务和 PVG 任务联合预训练, 提出了 GLIP. GLIP 设计并使用了一个基于对比学习思想的短语-区域对齐矩阵, 可以从目标检测数据中进行学习, 进而提升模型在处理 PVG 任务时的性能.

尽管上述工作在 PVG 任务上取得了不错的进展, 但都没有关注到短语和区域间的隐式对齐关系问题, 而预测这种关系可以有效评估模型深层多模态语义的理解能力. 本文考虑了 PVG 任务中的隐式对齐关系问题, 并标注了一个隐式数据集用于评估模型深层多模态语义的理解能力.

1.2 因果推理

最近, 因果推理在场景图生成[30]、语义分割[31]、视觉问答任务[32]等多个领域应用广泛, 引起了研究者极大的关注. Pearl 等人[9]将事物间的关系定义为 3 个层次: 关联、干预和反事实. 与传统的关联学习相比, 因果推理在减轻伪相关性和解耦模型效应以实现更好的泛化性方面有很好的潜力. 本文主要关注于使用因果推理中的干预策略[33,34], 以缓解模型在建模隐式关系时会被浅层语义所混淆的问题.

后门调整和前门调整是干预中两种常用的策略[9], 用以解决潜在的混杂因子的问题, 从而进一步解决关联学习中的伪相关问题. 对于后门调整策略, Wang 等人[35]认为当训练和测试数据是独立同分布时, 混杂因素会欺骗注意力机制来捕捉数据中有利于预测的伪相关性. 他们提出了一个基于后门调整策略的因果注意模块, 以无监督的形式对混杂因素进行自我注释来缓解混杂因子的影响. Huang 等人[36]认为在视觉推理任务中, 混杂偏差是制约任务性能的主要瓶颈, 并利用后门调整策略设计了一个参考表达去混杂方法来消除混杂偏差. 对于前门调整策略, Yang 等人[37]使用前门调整策略设计了一种新的注意力机制 CATT, 用以消除现有基于注意力的视觉语言模型中不断变化的混杂效应. CATT 遵循了传统注意力机制的 Q-K-V 设定, 可以替换 Transformer 中任意的注意力模块.

受上述已有工作的启发, 本文将因果推理引入 PVG 任务中, 提出了一种新的隐式增强的因果建模短语视觉定位方法. 该方法基于前门调整的策略, 设计了一个隐式感知的因果注意力模块 (ICA) 来缓解模型在预测隐式短语对应区域时, 容易被其他浅层语义所混淆的问题.

2 PVG 任务的因果图

在介绍本文所提出的隐式增强的因果建模短语视觉定位方法前, 本文先概述了针对 PVG 任务构建的因果图. 在本节中, 首先简要介绍因果图中的各个变量以及它们之间的因果关系 (第 2.1 节); 然后介绍因果图中的前门路径和后门路径, 并使用前门调整策略实现隐式感知的因果干预的过程 (第 2.2 节).

2.1 PVG 任务因果图的构建

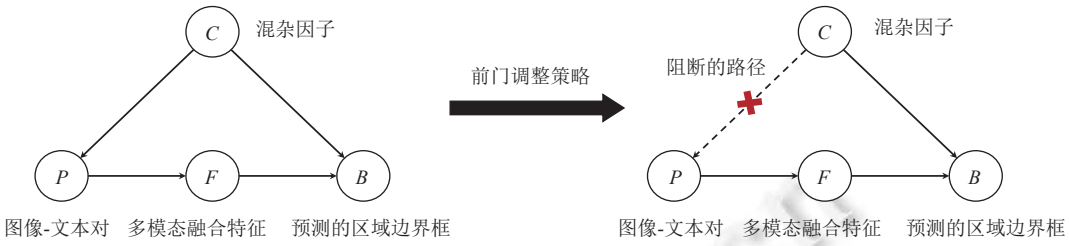
本文针对 PVG 任务构建的因果图如图 2 所示, 其中 P 代表图像-文本对, F 代表多模态融合特征, B 代表短语-区域预测边界框, C 代表混杂因子. 本文中 p, f, b 分别代表 P, F, B 的观测值.

$P\rightarrow F\rightarrow B$ 表示从图像-文本对 P 到短语-区域预测边界框 B 的预期因果效应, 其中多模态融合特征 F 起到中介作用. 在传统的 PVG 方法中, 模型首先分别提取图像和文本特征, 接着将图像特征和文本特征融合得到多模态融合特征表示, 最后利用多模态融合特征预测文本中的短语在图像中的区域边界框.

$P\leftarrow C\rightarrow B$ 表示不可见的混杂因素 C 对图像-文本对 P 和短语-区域预测边界框 B 的因果效应. 例如, 在图 1(c), (d) 中, “支持”是具有深层语义的短语, 模型需要理解“举着手”的动作是一种表示“支持”的常识才能将其对齐到“人们举着手”的区域. 而预测这种具有深层语义的短语所对应的区域时, 模型往往会被其他一些浅层语义所混淆. 如:

© 中国科学院软件研究所 http://www.jos.org.cn

“举着手”是人的一种特有动作, 模型可能会错误地将具有浅层语义的样本 (即短语“一男一女 (man and woman)”和“一些人 (some people)”及其所在的区域) 与短语“支持”联系起来, 导致预测结果出现偏差. 本文中, 我们把这种会干扰模型预测结果的因素称为混杂因素, 并且这种混杂因素是不可见的.



(a) 传统PVG任务的因果图 (b) 消除混杂因子影响后的因果图

图 2 针对 PVG 任务构建的因果图

2.2 PVG 的前门调整策略

如果两个不相关的变量 P 和 B 在某个共因 C 的作用下产生了伪相关, 则共因 C 被称为“混杂因子”, 在因果图中, 从 C 出发连接 P 和 B 的路径 $P\leftarrow C\rightarrow B$ 称为 P 和 B 之间的“后门路径”. 相应地, $P\rightarrow B$ 即为“前门路径”. 若以 C 为条件可以阻断 P 和 B 之间所有的后门路径 (即 P 和 B 之间无法通过 C 联系起来), 并且 C 的全部数据可以通过观测得到, 则这种调整策略被称为后门调整策略. 反之, 若 C 的数据无法收集, 同时 P 和 B 之间存在一个中介 F 满足: 1) F 阻断了所有 P 到 B 之间的路径; 2) P 到 F 之间没有未被阻断的路径; 3) F 到 B 之间的所有后门路径被 P 阻断, 则 F 满足 P 和 B 之间的前门准则. 此时, 以 P 为条件, F 为中介因素进行干预调整的策略被称为前门调整策略. 根据第 2.1 节中的描述, 本文中的混杂因子 C 不可见, 同时 P 和 B 之间存在一个中介 F 满足前门准则. 因此, 本文采用前门调整策略, 以 P 为条件来缓解模型在建模隐式关系时会被浅层语义所混淆的问题, 如图 2(b) 所示.

给定图像-文本对 P, PVG 任务的因果推理过程是根据文本短语最大化图像中与之相对齐的区域边界框 B 的干预条件概率, 而这一概率并不等于观察条件概率, 即:

$$\Pr(B=b\mid do(P=p))\neq\Pr(B=b\mid P=p) \tag{1}$$

其中, do 操作表示通过干预 $P=p$ 来实现 P 对 B 的因果效应, 在因果图中, do 操作将删除所有指向该变量的边. 如图 2(b) 中的虚线所示, 在以 P 为条件时, 我们阻断了后门路径 $F\leftarrow P\leftarrow C\rightarrow B$, 从而消除了混杂因子 C 对图像文本对 P 的影响, 进而我们可以使用前门调整策略计算 $P\rightarrow F\rightarrow B$ 的因果效应:

$$\Pr(B=b\mid do(P=p))=\sum_{f}\Pr(F=f\mid P=p)\sum_{p}\Pr(P=p)[\Pr(B=b\mid P=p,F=f)] \tag{2}$$

3 隐式增强的因果建模短语视觉定位方法 (ICM)

本节将会详细介绍隐式增强的因果建模短语视觉定位方法 (ICM). 首先简要说明 PVG 的任务定义 (第 3.1 节); 然后描述图像和文本模态的特征编码过程 (第 3.2 节); 接着详细介绍隐式感知的因果注意力模块 ICA (第 3.3 节); 最后介绍针对 PVG 任务设计的隐式感知的优化模块 (第 3.4 节).

3.1 任务定义

给定一组包含 T 个图像-文本对的集合 (V,S), 其中 $V=[V_1,\ldots,V_i,\ldots,V_T]$, $S=[S_1,\ldots,S_i,\ldots,S_T]$. 每个文本 S_i 包含多个短语 s_i^p, 即 $[s_i^1,\ldots,s_i^p,\ldots,s_i^n]\in S_i$, 其中 n 代表文本中的短语个数. 每个短语 s_i^p 都指向图像 V_i 中的一个或几个区域 $[r_i^1,\ldots,r_i^q]$, 其中 q 代表该短语对应的区域数量. 如图 3 所示, 对给定文本 S_i, PVG 任务的目的是准确预测其包含的每个短语 s_i^p 对应图像 V_i 中的区域的边界框 $b_i^p=x,y,w,h$, 其中 (x,y) 代表预测区域边界框的中心点坐标, w 和 h 分别代表预测区域边界框的宽度和高度.

© 中国科学院软件研究所 http://www.jos.org.cn

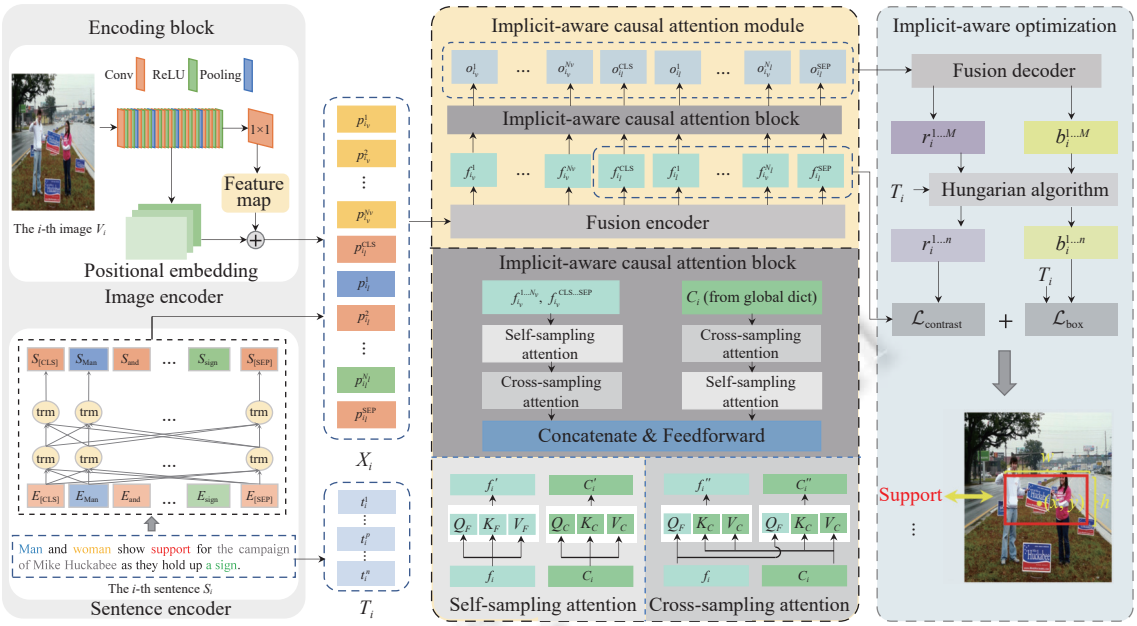


图 3 隐式增强的因果建模短语视觉定位方法 (ICM) 整体网络结构图

3.2 图像-文本特征编码模块

(1) 图像特征编码

对给定的图像 $V_i \in \mathbb{R}^{3\times H_0\times W_0}$, 本文使用在计算机视觉领域有广泛应用的 ResNet 神经网络[38]作为基础视觉模型来提取每张图像 V_i 的 2D 特征图 $V_0 \in \mathbb{R}^{D\times H\times W}$, 其中通道维度 D 为 2 048, 特征图的高度 H 和宽度 W 分别为初始高度 H_0 和初始宽度为 W_0 的 1/32. 然后, 本文使用一个 1×1 的卷积层将通道维度 D 降为 D_v= 256. 接着, 将特征图扁平化得到最终的特征图 $V_i \in \mathbb{R}^{Dv\times Nv}$, 其中 $N_v = H\times W$. 为了保存原始 2D 特征图的空间信息, 本文遵循 MDETR[7]的设定, 将位置编码添加到 V_i 中, 如图 3 所示.

(2) 文本特征编码

传统的文本特征是通过 GloVe embedding[39]对文本中的每个单词进行表示. 随着预训练语言模型在 NLP 领域的大规模应用, 本文采用预训练的 RoBERTa-base[40]模型来提取文本特征. 相较于使用 GloVe 获得的 300 维的文本特征向量, 本文采用 RoBERTa-base 模型得到 768 维的文本特征向量 $L_i \in \mathbb{R}^{D_l\times(N_l+2)}$, 其中 D_l = 768, N_l 为文本长度, 2 为在编码文本时在文本的头部和尾部添加的两个 token [CLS] 和 [SEP] 的长度. 在文本特征提取的过程中, 本文根据数据集中已标注的信息, 记录下文本中短语的起始位置以及该短语在图像中所对应的区域边界框, 记为 t_p, 最终得到 T_i=[t_1,…,t_p,…,t_n], 其中 n 为文本中短语的个数, 如图 3 所示.

在得到图像特征和文本特征后, 我们采用两个全连接层将其映射到同一个特征空间 $\mathbb{R}^D$ 中, 投影后的视觉特征和文本特征分别为 p_{i_v}, p_{i_l}. 将 p_{i_v} 和 p_{i_l} 进行拼接得到多模态的特征表示 X_i:

$$X_i = [\underbrace{p_{i_v}^1, p_{i_v}^2, \ldots, p_{i_v}^{N_v}}_{\text{视觉特征 } p_v}, \underbrace{p_{i_l}^{\text{CLS}}, p_{i_l}^1, \ldots, p_{i_l}^{N_l}, p_{i_l}^{\text{SEP}}}_{\text{文本特征 } p_l}] \tag{3}$$

3.3 隐式感知的因果注意力模块 (ICA)

图 1(c), (d) 中, 相较于“标语 (a sign)”“彩虹旗 (rainbow flags)”这些具有浅层语义的短语, “支持 (support)”这类短语的语义较深并且稀疏, 需要模型进一步理解深层语义才能将其与“人们举着手的动作”这一常识性区域相对齐. 虽然已有的 PVG 方法取得了不错的进展, 但是它们普遍忽略了类似“支持 (support)”的隐式短语-区域对齐关系问题. 基于此, 本文提出了一种隐式增强的因果建模短语视觉定位方法 (ICM) 来有效建模短语-区域间的隐式关系.

© 中国科学院软件研究所 http://www.jos.org.cn

其中, ICM 的核心组成部分为隐式感知的因果注意力模块 (ICA), 用以缓解模型在建模隐式关系时会被浅层语义所混淆的问题.

ICA 的核心思想如图 3 中的 implicit-aware causal attention module 所示. 在得到公式 (3) 的多模态的特征表示 $X_i \in \mathbb{R}^{D\times(N_v+N_l+2)}$ 后, 本文首先将其输入到 fusion encoder 模块中得到多模态融合特征 F_i:

$$F_i = [\underbrace{f_{i_v}^1, f_{i_v}^2, \ldots, f_{i_v}^{N_v}}_{\text{视觉特征 } f_v}, \underbrace{f_{i_l}^{\text{CLS}}, f_{i_l}^1, \ldots, f_{i_l}^{N_l}, f_{i_l}^{\text{SEP}}}_{\text{文本特征 } f_l}] \tag{4}$$

已有的 PVG 方法大多基于 F_i 直接进行区域边界框的预测, 但是 F_i 中的特征信息并没有针对隐式关系进行有效建模, 导致模型在预测具有隐式信息的短语所对应的区域边界框时结果会出现偏差. 本文提出的 ICM 方法基于公式 (2) 的前门调整策略, 并使用注意力机制来对公式 (4) 得到的多模态融合特征 F_i 实施该策略来消除混杂偏差, 从而有效建模隐式关系. 考虑到对所有的样本全部进行前向传播高昂的计算代价, 本文对 P 和 F 进行采样, 并将其输入模型中来计算 $P(B=b\mid do(P=p))$. 此外, 本文引入了归一化加权几何平均 (NWGM)[41,42]近似来实现公式 (2) 的目标:

$$P(B=b\mid do(P=p)) \approx Softmax[g(\hat{F},\hat{P})] \tag{5}$$

$$\hat{F} = \sum_f P(F=f\mid h(P))\boldsymbol{f} \tag{6}$$

$$\hat{P} = \sum_p P(P=p\mid j(P))\boldsymbol{p} \tag{7}$$

其中, $g(\cdot)$ 是用于公式 (2) 中分布 $P(B=b\mid P=p,F=f)$ 的参数化网络, 并使用 $Softmax$ 将其归一化. 此外, $\hat{F}$ 和 $\hat{P}$ 分别代表自采样 (self-sampling) 和交叉采样 (cross-sampling) 的估算值, $\boldsymbol{f}$ 和 $\boldsymbol{p}$ 是对应于变量 f 和 p 的嵌入式向量. 函数 $h(\cdot)$ 和 $j(\cdot)$ 用来将输入的 P 转换为两个不同的可以被参数化为网络的查询集合.

实际上, $\hat{F}$ 和 $\hat{P}$ 是经典的注意力机制所计算的内容, 可以通过使用 Q-K-V 操作简单地表示为图 3 中的 self-sampling attention 和 cross-sampling attention 模块, 因此, 自采样 $\hat{F}$ 和交叉采样 $\hat{P}$ 可以使用如下的公式表示:

$$\hat{F} = \begin{cases} V_F Softmax(Q_F^{\top}K_F) & \text{(a)} \\ V_C Softmax(Q_C^{\top}K_C) & \text{(b)} \end{cases} \tag{8}$$

$$\hat{P} = V_C Softmax(Q_F^{\top}K_C) \tag{9}$$

公式 (8) 和公式 (9) 分别代表自采样和交叉采样. 公式 (8) 的 (a) 计算了多模态融合特征 F 的 self-sampling attention, (b) 计算了混杂因子 C 的 self-sampling attention. 在具体实现中, Q_F 来自 $h(P)$, Q_C 来自 $j(P)$; K_F 和 V_F 来自当前输入的样本; K_C 和 V_C 来自训练集中的其他样本, 并作为从整个训练集压缩而来的全局词典. 具体而言, 我们通过对训练集所有样本的嵌入 (如图像的 RoI 特征) 进行 K-means 聚类操作[43]来初始化这个词典.

基于公式 (5), (8) 和 (9), 我们可以实现公式 (2) 中的前门调整策略, 从而计算出 $P\rightarrow F\rightarrow B$ 的因果效应, 得到的输出如下所示:

$$O_i = [\underbrace{o_{i_v}^1, o_{i_v}^2, \ldots, o_{i_v}^{N_v}}_{\text{视觉特征 } o_v}, \underbrace{o_{i_l}^{\text{CLS}}, o_{i_l}^1, \ldots, o_{i_l}^{N_l}, o_{i_l}^{\text{SEP}}}_{\text{文本特征 } o_l}] \tag{10}$$

O_i 的维度和 F_i 完全一致, 可以直接进行区域边界框的预测. 相较于 F_i, O_i 对隐式关系进行了有效建模, 用以帮助模型准确预测图像中与具有深层语义的隐式短语相对齐的区域. 同时, 通过前门调整策略, O_i 消除了 F_i 中的浅层语义所带来的混淆信息. 通过这种方式, 相较于已有的 PVG 方法, ICM 可以有效建模隐式短语-区域对齐关系, 提升了模型多模态深层语义理解的能力.

3.4 隐式感知的优化模块

在进行边界框预测时, ICM 首先将 O_i 输入图 3 中的 fusion decoder 中, 得到一个包含 M 个区域和 M 个边界框的预测集合:

© 中国科学院软件研究所 http://www.jos.org.cn

$$\begin{cases} R_i = [r_i^1, \ldots, r_i^s, \ldots, r_i^M] \\ B_i = [b_i^1, \ldots, b_i^s, \ldots, b_i^M] \end{cases} \tag{11}$$

其中, r_i^s 为图像中的区域, b_i^s=x,y,w,h 为该区域所对应的 4 维边界框.

对文本 S_i 中的每个短语 s_i^p, ICM 基于第 3.2 节中文本特征抽取步骤记录下的文本短语信息 T_i, 利用匈牙利算法 (Hungarian algorithm)[44]从集合 R_i 和 B_i 选出与之最对齐的预测区域 r_i^p 和预测边界框 b_i^p. PVG 任务的目标是准确预测每个短语所对应的边界框, 因此, 本文的优化目标是最大化预测边界框 b_i^p 与真实边界框 g_i^p 的交并比 (IoU). 本文首先采用 TransVG 等方法[20,27]中常用的边界框损失, 记为 $\mathcal{L}_{\text{box}}$, 用以最大化 b_i^p 和 g_i^p 的重叠面积, 从而最大化 IoU; 此外, 受 GLIP[8]的启发, 本文额外引入了一个对比对齐损失 $\mathcal{L}_{\text{contrast}}$, 用以确保 s_i^p 与集合 R_i 中与 s_i^p 相对齐的区域在特征空间中比不对齐的区域更接近. 具体而言, 对于短语 s_i^p, 该损失同样使用了 s_i^p 在文本中的起始位置, 用以提取公式 (4) 中 F_i 的文本特征 f_l 对应位置的短语特征信息, 并使用该短语特征信息对齐目标检测所预测的区域, 使得匈牙利算法对齐出的预测边界框更接近真实边界框.

对于 $\mathcal{L}_{\text{box}}$, 本文使用 *GIoU* 损失[45]和 L1 损失来实现对预测结果的优化:

$$\mathcal{L}_{\text{box}} = \sum_{i=1}^{n} \lambda_{\text{giou}} \mathcal{L}_{\text{giou}}\left(b^i, \hat{b}^i\right) + \lambda_{\text{L1}} \mathcal{L}_{\text{L1}}\left(b^i, \hat{b}^i\right) \tag{12}$$

$$\mathcal{L}_{\text{giou}}\left(b^i, \hat{b}^i\right) = 1 - GIoU\left(b^i, \hat{b}^i\right) \tag{13}$$

$$\mathcal{L}_{\text{L1}}\left(b^i, \hat{b}^i\right) = \left\| b^i - \hat{b}^i \right\|_1 \tag{14}$$

在公式 (12) 中, n 为文本中的短语数量, b^i 为短语的预测边界框, $\hat{b}^i$ 为短语所对应的真实边界框. λ_{giou} 和 λ_{L1} 分别是用以平衡 *GIoU* 损失和 L1 损失的超参数.

对于 $\mathcal{L}_{\text{contrast}}$, 本文将使用匈牙利算法选出的与待预测短语最对齐的预测区域作为对比学习中的正样本, 其余 N–1 个预测区域作为负样本, 如下所示:

$$\mathcal{L}_{\text{contrast}} = -\sum_{i=1}^{n} \sum_{j=1}^{N} \log \frac{\exp(s^i \cdot r_+^i / \tau)}{\exp(s^i \cdot r^j / \tau)} \tag{15}$$

其中, n 为文本中的短语数量, N 为预测区域的数量, s^i 为当前待预测短语, r_+^i 为与 s^i 对齐的预测区域, r^j 为预测区域集合中的区域, τ 为对比学习中的温度超参数.

本文使用上述两个损失函数联合优化 ICM 中的参数, 并引入超参数 $\lambda_{\text{contrast}}$ 来平衡 $\mathcal{L}_{\text{contrast}}$ 的权重, 最终的损失函数 $\mathcal{L}_{\text{total}}$ 如公式 (16) 所示:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{box}} + \lambda_{\text{contrast}} \mathcal{L}_{\text{contrast}} \tag{16}$$

4 实 验

本节描述了实验细节, 包括针对隐式场景构建的隐式数据集 (第 4.1 节), Baselines 方法 (第 4.2 节), 实验设置及评价指标 (第 4.3 节), 实验结果对比 (第 4.4 节) 以及实验分析 (第 4.5 节).

4.1 隐式数据集构建

本文首先基于对 Flickr30k Entities 数据集[46]的分析, 通过预标注总结了如图 4 所示的 4 种主要隐式关系, 并构建了一个面向隐式场景的数据集. 其中, 4 种隐式关系定义如下.

- 常识性理解表示模型需要理解短语中蕴含的具有常识性信息的深层语义, 例如: 为准确预测“支持 (support)”所指向的区域, 模型需要理解“举着手”的动作是一种表示“支持”的常识.
- 上下文理解表示模型需要从文本的上下文信息中理解短语蕴含的深层语义信息, 例如: 在没有上下文信息的情况下, “另外三个人 (three more people)”很难准确与“图像左下角的三个人”这一区域对应.
- 空间关系理解表示模型需要对空间关系进行有效建模, 例如: “在另一个男人的旁边 (next to another man)”包含了两个男人之间的位置信息.

© 中国科学院软件研究所 http://www.jos.org.cn

● 数值信息理解表示一个短语可能指向的是多个区域的情况, 需要模型准确理解数值类信息进而精确预测, 例如: “其中三位 (three of them)”指向了图像中的 3 个区域.

隐式关系类别	占比 (%)	带有隐式关系的图像-文本对
常识性理解	34.5	Man and woman show **support** for the campaign of Mike Huckabee as they hold up a sign. 一男一女高举标语, 表示支持迈克-赫卡比的竞选活动.
上下文理解	26.9	Two people feeding sheep in a field with a dog nearby and **three more people** looking at them. 两个人在田野里喂羊, 旁边有一只狗, 还有三个人看着他们.
空间关系理解	23.5	A man is playing a guitar **next to another man** who is sitting behind a green cart wearing a mask. 一名男子正在弹吉他, 旁边还有一名男子戴着面具坐在一辆绿色小车后面.
数值信息理解	15.1	Six ladies at the dining table and **three of them** are knitting. 餐桌上坐着六位女士, 其中三位正在编织.

图 4 面向隐式场景的 4 种主要短语-区域对齐关系以及每种隐式关系在隐式数据集中的占比

在数据标注过程中, 本文邀请了两位标注人员同时对数据集中的短语-区域对进行隐式或非隐式的标注, 在标注过程中, 若两位标注人员无法就某一短语-区域对是否为隐式关系达成一致, 我们将另外安排一位专家来做最终的决定. 标注结束后进行 Kappa 一致性检测, 最终 Kappa 检测值为 0.85, 说明了此次数据标注的一致性. 由于标注非常复杂, 费时且困难, 因此本文基于如图 4 所示的 4 种主要隐式关系对 Flickr30k Entities 原始数据集的测试集和验证集共 15k 条短语-区域对信息进行了标注. 最终, 我们得到了 1.4k 条隐式短语-区域对信息, 12.73k 条显式短语-区域对信息. 此外, 原始数据集中还存在一定量 (0.87k 条) 的错误标注与冗余标注的数据. 例如, 原始数据集中, 图 1(c) 的“支持 (support)”短语除了与“人举着手”的区域对应外, 还与一些人所在的区域对应. 本文的隐式与显式数据集可以看作是对原始数据集进一步的精标注, 过滤了错误与冗余的部分. 因此隐式数据集加上显式数据集实际上是原始数据集的子集. 按照数据集的原始划分, 我们得到了隐式数据集和显式数据集的验证集和测试集.

在训练阶段, 本文使用包含隐式和显式短语-区域对齐关系的全部数据集进行训练 (即 Flickr30k Entities 数据集的原始训练集); 在验证和测试阶段, 本文使用隐式数据集, 显式数据集以及 Flickr30k Entities 原始的数据集进行验证和测试, 分别得到验证集和测试集上的实验结果.

4.2 Baselines

本文通过对比传统的性能优异的 PVG 方法, 多模态预训练模型, 多模态大语言模型 (MLLM) 来验证 ICM 的有效性, 如下所述.

(1) 传统的 PVG 方法

● FAOA (a fast and accurate one-stage approach)[11]较早提出了单阶段方法, 使用 YOLOv3 作为目标检测器来提取图像特征, 并将文本特征融合进 YOLOv3 中. 此外, 考虑到图像中的空间信息, 视觉文本融合特征还采用空间特征进行特征数据增强.

● ReSC (recursive sub-query construction framework)[23]提出了一种递归子查询构建框架来解决 PVG 方法在面对长且复杂的文本查询时定位效果不佳的问题. 其设计使用了一种子查询学习器来构建子查询, 该学习器使用子查询调制网络来利用子查询完善视觉文本特征, 利用子查询最后一轮的视觉文本特征预测最终的区域边界框.

● TransVG (Transformers for visual grounding)[27]是一种基于 Transformer 编码器的堆叠结构设计的视觉定位模型, 它解决了传统的两阶段和单阶段方法对于手工设计推理模块和多模态融合模块 (如: 图像场景图等) 的过度依赖, 从而导致模型容易过度拟合具有特定场景的数据集的问题. 此外, TransVG 将视觉定位问题定义为直接坐标回归问题, 直接进行区域边界框的预测, 避免了从一组候选区域进行预测.

© 中国科学院软件研究所 http://www.jos.org.cn

● VLTVG (visual-linguistic verification for visual grounding)[19]也是一种基于 Transformer 设计的视觉定位方法, 它设计了一种语言引导的视觉特征聚合方法和多级跨模态解码器, 用以关注图像中与文本描述相关的特征, 同时抑制与文本不相关的区域特征, 从而提供视觉特征的显著性. 然而, 由于忽略了多层次的模态信息, 其性能仍有待提高.

● QRNet (query-modulated refinement network)[20]和 VLTVG 同样希望更多地关注到视觉特征信息. QRNet 认为视觉分支模型提取的图像特征和多模态推理真正需要的特征是不一致的. 因此, 它通过一个新颖的查询感知动态注意力 (QD-ATT) 机制和查询感知的多尺度融合策略来调整视觉模型的中间特征, 从而解决不一致问题. 在视觉模型生成的图像特征图的空间和通道级别中, QD-ATT 可动态计算依赖于文本查询的视觉注意力.

(2) 全监督多模态预训练模型

文本的预训练推动了多模态预训练的发展. 通过在大规模多模态数据上进行预训练, 模型捕捉文本和视觉信息之间的语义关系的能力得以提升, 从而提高了其在多种多模态下游任务上的性能. 本文选取了两种在 PVG 下游任务上有优异性能的模型来验证 ICM 方法的潜力.

● MDETR (modulated detector)[7]是一种基于 DETR[29]目标检测网络的预训练模型, 它将 PVG 任务建模为一个调制检测任务, 使用 130 万个图像文本对 (包含 Flickr30k Entities) 进行预训练, 而这些图像文本对来自现有的多模态数据集, 且文本中的短语与图像中的区域有明确的对应关系.

● GLIP (grounded language-image pre-training)[8]将目标检测任务和 PVG 任务联合训练, 使用了比 MDETR 更多的 2 700 万个图像文本对 (包含 Flickr30k Entities) 进行预训练, 其中包括 300 万个人工标注的高质量图像文本对和 2 400 万个网络抓取的图像文本对.

本文在隐式和显式数据集的测试集上评估 MDETR, GLIP 和 ICM 三者的性能, 并将结果绘制成如图 5 所示的训练数据量-性能散点图, 其中, M 代表百万个图像-文本对.

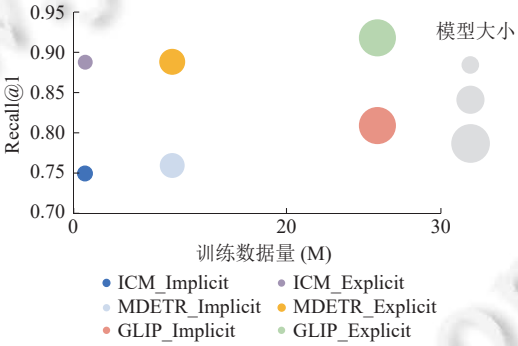


图 5 ICM 与 MDETR 和 GLIP 在训练数据量及在隐式和显式数据集的性能对比

(3) 自监督多模态大语言模型 (MLLM)

最近, 随着 ChatGPT 等大语言模型 (LLM) 的兴起, 多模态大语言模型 (MLLM) 也有了快速的发展[47,48] (如 GPT-4 等), 它们已经在多模态任务的多个领域展现出优异的 zero-shot 性能. 由于 GPT-4 并没有开放源代码, 本文选取了 2 个开源 MLLMs 来验证 ICM 方法的有效性.

● MiniGPT4-13B[47]利用视觉编码器 BLIP-2[49]和大语言模型 Vicuna 进行训练, 使用一个投影层将来自 BLIP-2 的视觉编码器和 Vicuna-13B 进行对齐. MiniGPT-4 采用两阶段训练方法: 第 1 阶段使用 500 万条图像文本对进行训练, 使得 Vicuna 初步具备理解图像的能力; 第 2 阶段使用 3 500 对高质量图像文本数据进行微调, 显著提升了 MiniGPT-4 的可靠性和可用性.

● LLaVA-13B[48]和 MiniGPT-4 的思想类似, 目的都是在于对齐视觉模型和大语言模型. LLaVA 选用 CLIP[6]作为其视觉基础模型, 采用和 MiniGPT-4 类似的两阶段方法进行训练. LLaVA 和 MiniGPT-4 的不同之处在于它不需要 MiniGPT-4 复杂的 Q-former 结构, 但是需要微调基础大语言模型.

© 中国科学院软件研究所 http://www.jos.org.cn

由于它们没有专门面向 PVG 任务测试的代码, 因此本文采用评测大模型常用的两种评价方式: zero-shot (ZS)[50] 和 in-context learning (ICL)[51]来评测它们在隐式和显式数据集上的性能, 并将结果绘制成如图 6 所示的柱状图.

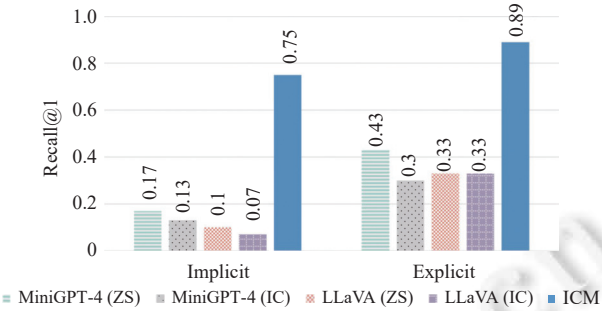


图 6 ICM 与 MiniGPT4-13B 和 LLaVA-13B 在隐式和显式数据集的性能对比

4.3 实验设置和评价指标

(1) 实验设置

实验中, 对于所有的 baselines, 本文使用其开源代码以及论文中汇报的超参数在标注的隐式数据集和显式数据集上进行实验, 并分别得到验证集和测试集的实验结果. 对于原始数据集, 我们引用其论文中汇报的实验结果.

对于 ICM, 本文使用 Adam 优化器进行梯度更新, 学习率设置为 1E–4, 自采样和交叉采样的层数设置为 6, 隐藏层的维度设置为 256, fusion decoder 得到的区域和边界框数量 M 为 100. 对于损失函数, 超参数 λ_{giou}, λ_{L1} 和 $\lambda_{\text{contrast}}$ 分别设置为 2, 5 和 1, $\lambda_{\text{contrast}}$ 的温度超参数 τ 设置为 0.07. 训练时, 批量大小设置为 8, 使用 2 张 40G A100 进行训练, 总训练 epoch 数为 60, 每轮 epoch 所需时间为 2 h. 同时, 为了防止过拟合, 本文采用丢弃率为 0.1 的 Dropout 策略.

(2) 评价指标

依据之前的工作[46], 本文采用召回率 (Recall) 作为评价指标. 对于一个查询短语, 若预测的边界框与实际真实的边界框的交并比 (IoU)≥0.5, 则认为对该短语所对应的区域预测正确, 此时的指标称为 Recall@1, 记为 R@1. 本文分别汇报了 ICM 方法与传统 PVG 方法 (如表 1 所示), 大规模全监督预训练方法 (如图 5 所示) 以及自监督多模态大语言模型 (如图 6 所示) 的性能对比.

表 1 ICM 方法与传统 PVG 方法在隐式数据集、显式数据集和原始数据集上的性能比较 (%)

Approach	隐式数据集 (implicit)		显式数据集 (explicit)		原始数据集 (full)	
	Val	Test	Val	Test	Val	Test
FAOA	64.94	61.32	72.35	71.31	—	68.69*
ReSC	66.44	62.58	76.75	77.64	—	69.28*
TransVG	70.03	69.08	80.92	82.65	—	79.1*
VLTVG	71.31	70.17	81.7	83.42	—	79.84*
QRNet	72.11	71.52	82.52	84.85	—	81.95*
ICM	**80.12**	**74.92**	**87.54**	**88.96**	**81.76**	**82.67**
ICM w/o ICA	76.64	71.62	85.45	86.27	78.43	79.83

注: 对于原始数据集, “*”表示该结果为引用原论文中汇报的结果, “—”表示原论文没有对该结果进行汇报

4.4 实验结果对比

表 1 展示了本文 ICM 方法和传统的 PVG 方法在隐式数据集, 显式数据集和原始数据集上的实验结果对比. 对于隐式数据集和显式数据集的结果, 本文使用各个方法的开源代码, 分别在标注的隐式数据集和显式数据集上进行实验. 分析表 1 的实验结果, 我们可以得到以下信息.

© 中国科学院软件研究所 http://www.jos.org.cn

(1) 建模隐式关系异常困难. 实验结果表明, 所有的方法在隐式数据集上的性能都显著低于在显式数据集上的性能. 例如, 对于传统 PVG 方法中性能最好的 QRNet, 隐式数据集比显式数据集 R@1 的平均结果降低了 11.87% (p-value<0.01); 对于 ICM 方法, 隐式数据集比显式数据集 R@1 的平均结果降低了 10.73% (p-value<0.01). 这表明相较于显式关系, 有效建模隐式关系是异常困难的.

(2) 对比所有传统的 PVG 方法, 本文的 ICM 方法在隐式数据集上的性能提升显著 (如对比现阶段 PVG 方法中性能最好的 QRNet, ICM 在验证集和测试集上 R@1 的平均结果提升了 5.71% (p-value<0.01)). 这表明相较于传统的 PVG 方法, 本文所提的 ICM 方法可以有效地建模隐式短语-区域关系. 同时, 在显式数据集上, 相较于 QRNet 方法, ICM 在验证集和测试集上 R@1 的平均结果提升了 4.57% (p-value<0.01); 在原始数据集的测试集上, 相较于 QRNet, ICM 的 R@1 结果提升了 0.72% (p-value<0.05). 这表明 ICM 既缓解了建模隐式关系时会被浅层语义混淆的问题, 又可以保持对显式关系的建模能力. 此外, 值得注意的是, ICM 在原始数据集上的提升相较于隐式和显式数据集较小. 本文分析主要是由于原始数据集中错误与冗余的数据造成的, 如第 4.1 节隐式数据集构建部分所述. 而 ICM 在本文标注的隐式与显式数据集上提升显著更能验证我们方法的有效性. 这一定程度上也反映了数据质量对于评价模型的公平性与重要性.

4.5 实验分析

(1) ICA 模块的有效性

本节设计了针对 ICM 方法中核心模块 ICA 的有效性分析实验, 进一步验证了因果干预中的前门调整策略在缓解浅层语义带来的混淆上的有效性. 从表 1 的实验结果可以看出: 相较于 ICM 的结果, ICM w/o ICA 的结果在隐式数据集, 显式数据集和原始数据集上 R@1 的平均结果分别下降了 3.3% (p-value<0.05), 2.39% (p-value<0.05) 和 3.09% (p-value<0.05). 这说明了在不使用前门调整策略时, ICM 对于隐式关系的建模能力会显著下降, 进一步验证了 ICA 模块可以有效建模隐式关系, 并鼓励我们使用因果干预策略来缓解混杂偏差.

(2) ICM 方法与大规模全监督预训练模型对比

图 5 展示了 ICM 方法与大规模全监督预训练模型 MDETR 和 GLIP 在训练数据量, 模型大小以及在隐式和显式数据集测试集上的结果对比. 从图 5 中可以看出, 本文 ICM 方法的训练数据量远小于 MDETR 和 GLIP, 其中 ICM 的训练数据量为 15 万条图像-文本对, MDETR 的训练数据量为 130 万条图像-文本对, GLIP 的训练数据量为 2 700 万条图像-文本对. 对比 ICM 与 MDETR 的结果可以发现, 即使 MDETR 使用了包含 Flickr30k Entities 在内的 130 万条训练数据, 本文的 ICM 方法在仅使用 15 万条训练数据的情况下, 可以在隐式和显式数据集上取得与 MDETR 非常接近的性能. 此外, GLIP 相较于 ICM 方法可以取得更优的性能. 这是合理的, 因为 GLIP 使用了包含 Flickr30k Entities 在内的 2 700 万条训练数据, 在 64 张 V100 的硬件条件下进行训练, 而本文的 ICM 方法仅使用了 15 万条训练数据在 2 张 A100 的硬件条件下训练. 上述对 MDETR, GLIP 以及本文 ICM 这 3 种方法的分析验证了 ICM 方法的潜力, 启发我们下一步工作可引入更多的数据并投入更多的计算资源提升方法的性能.

(3) ICM 方法与自监督预训练多模态大语言模型对比

图 6 展示了 MiniGPT-4 和 LLaVA 两个自监督预训练的多模态大语言模型在隐式和显式数据集测试集上的 zero-shot 和 in-context learning 结果以及 ICM 方法的 R@1 结果对比. 由于 MiniGPT-4 和 LLaVA 并没有针对 PVG 任务进行测试的代码, 所以本文在测试集中随机选取了 30 对含有隐式和显式关系的图像-文本对, 采用大模型常用的两种性能评价方法进行测试并汇报 R@1 的结果: 1) zero-shot (ZS). 遵循 Bang 等人[50]提出的评估大模型的 zero-shot 设置, 本文给定一段任务定义以及图像-文本对, 对给定的短语, 我们要求 MLLM “generate a bounding box for the given phrase”. 2) in-context learning (ICL). 遵循 Dong 等人[51]提出的评估大模型的 ICL 设置, 我们在 ZS 设定的基础上给定几个短语-区域的例子“phrase, box;…; phrase, box”作为提示, 同样让模型生成给定短语对应的区域框. 对于 ZS 和 ICL 生成的区域框, 我们将其与图像中标注的区域框进行 IoU 计算, 得到 R@1 的结果. 如图 6 所示, 相较于 MiniGPT-4 和 LLaVA, ICM 的性能均远超它们. 这表明现有的基于图像-文本对训练的多模态大语言模型在文本上缺乏理解深层语义的能力, 在图像上缺乏理解细粒度图像的能力. 此外, 我们还发现, ZS 的性能高

© 中国科学院软件研究所 http://www.jos.org.cn

于 ICL 的性能, 原因在于 MiniGPT-4 和 LLaVA 分别使用 BLIP-2[49]和 CLIP[6]作为图像编码模型, 而论文中已经说明它们并没有 ICL 的能力[6,49].

(4) 隐式案例分析

为了更直观地验证本文 ICM 方法在预测隐式短语-区域对齐关系方面的有效性, 本文从标注的隐式数据集的 4 种隐式关系中各随机选取了 1 个样本进行分析, 如图 7 所示. 图 7 中, (a1), (b1), (c1), (d1) 代表隐式短语所对应区域的真实边界框 (图像中的红色框); (a2), (b2), (c2), (d2) 代表 QRNet 方法的预测结果 (图像中的黄色框); (a3), (b3), (c3), (d3) 代表 ICM 方法的预测结果 (图像中的白色框). 从图 7 中可以发现, 在面对含有隐式关系的短语-区域对时, QRNet 预测的结果相较于短语所对应的真实区域存在较大偏差, 而 ICM 的预测结果通常可以正确对齐隐式短语所对应的图像区域. 图 7(a) 中, QRNet 对“支持”的预测为包含人的区域, 而 ICM 的预测为仅包含人举着标语的区域; 图 7(b) 中, QRNet 对“还有三个人”的预测区域为中心两个人喂羊的区域, 而 ICM 可以准确预测该短语对应区域为图像左下角的旁边 3 个人所在区域; 图 7(c) 中, QRNet 对“旁边还有一名男子”的预测区域为两人中间的区域, ICM 则可以做出正确的预测; 图 7(d) 中, QRNet 对“其中三位”的预测区域为图像中随机的 3 个区域, 并不是正在编织的 3 个人, ICM 的预测区域则为正在编织的 3 个人所对应的正确区域. 这再次验证了本文所提的 ICM 方法在预测隐式短语-区域对齐关系上的有效性.

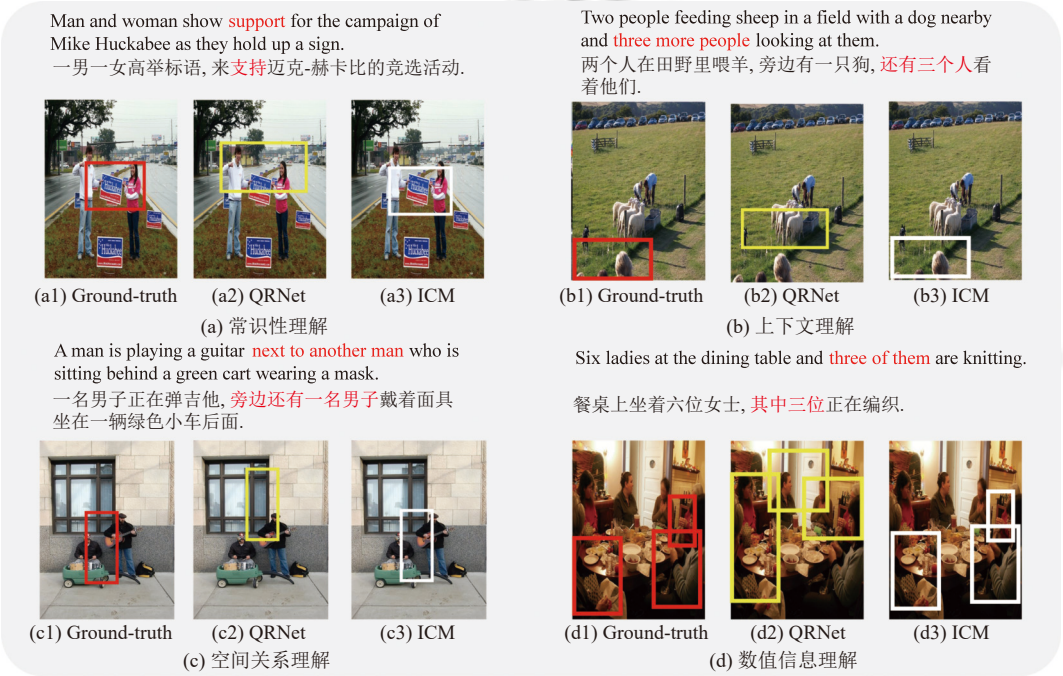


图 7 QRNet 和 ICM 对 4 种隐式关系的预测结果对比

5 总 结

本文针对短语视觉定位 (PVG) 任务中的隐式短语-区域对齐关系进行了研究. 通过分析已有的 PVG 数据集, 本文发现了其中存在的隐式短语-区域对齐关系问题, 并总结了 4 种隐式关系, 构建了一个面向隐式场景的数据集. 然而有效建模这种隐式关系异常困难, 已有的传统 PVG 方法无论是两阶段方法还是单阶段方法, 都将重心放在如何学习文本短语和图片区域之间的关联上, 忽视了短语和区域间的隐式对齐关系问题, 预测结果常会被一些浅层语义所混淆. 本文分析认为, 有效建模隐式关系需要引入因果干预方法来缓解浅层语义带来的混淆问题, 并为 PVG 任务提出了一种隐式增强的因果建模短语视觉定位方法 ICM, 其通过使用因果推理中的前门调整策略来有效地

© 中国科学院软件研究所 http://www.jos.org.cn

建模隐式关系. 在隐式数据集上的实验结果表明, ICM 的性能优于现有的传统 PVG 方法, 验证了其在建模隐式关系方面的有效性. 此外, ICM 的性能对比一些多模态大语言模型仍有优势, 说明已有的多模态大语言模型暂且没有对隐式关系有效建模. 在未来的工作中, 我们计划引入更多的信息 (如知识图谱) 来帮助对齐隐式短语-区域. 此外, 我们计划将 ICM 方法迁移到其他也存在隐式关系的任务中, 如目标指代理解 (REC) 和视频定位 (video grounding).

致谢 本文工作受软件新技术与产业化协同创新中心部分资助.

References:

[1] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

[2] Wang LW, Li Y, Huang J, Lazebnik S. Learning two-branch neural networks for image-text matching tasks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(2): 394–407. [doi: 10.1109/TPAMI.2018.2797921]

[3] Hossain MDZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR), 2019, 51(6): 118. [doi: 10.1145/3295748]

[4] Datta R, Joshi D, Li J, Wang JZ. Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR), 2008, 40(2): 5. [doi: 10.1145/1348246.1348248]

[5] Antol S, Agrawal A, Lu JS, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D. VQA: Visual question answering. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2425–2433. [doi: 10.1109/ICCV.2015.279]

[6] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[7] Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I, Carion N. MDETR—Modulated detection for end-to-end multi-modal understanding. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 1760–1770. [doi: 10.1109/ICCV48922.2021.00180]

[8] Li LH, Zhang PC, Zhang HT, Yang JW, Li CY, Zhong YW. Grounded language-image pre-training. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 10955–10965. [doi: 10.1109/CVPR52688.2022.01069]

[9] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. New York: Basic Books, Inc., 2018.

[10] Kazemzadeh S, Ordonez V, Matten M, Berg T. Referitgame: Referring to objects in photographs of natural scenes. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 787–798. [doi: 10.3115/v1/D14-1086]

[11] Yang ZY, Gong BQ, Wang LW, Huang WB, Yu D, Luo JB. A fast and accurate one-stage approach to visual grounding. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 4682–4692. [doi: 10.1109/ICCV.2019.00478]

[12] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

[13] Yu LC, Lin Z, Shen XH, Yang JM, Lu X, Bansal M, Berg TL. MattNet: Modular attention network for referring expression comprehension. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1307–1315. [doi: 10.1109/CVPR.2018.00142]

[14] Zhuang BH, Wu Q, Shen CH, Reid I, van den Hengel A. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4252–4261. [doi: 10.1109/CVPR.2018.00447]

[15] Yu Z, Yu J, Xiang C, Xiang CC, Zhao Z, Tian Q, Tao DC. Rethinking diversified and discriminative proposal generation for visual grounding. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: AAAI Press, 2018. 1114–1120.

[16] Yang SB, Li GB, Yu YZ. Dynamic graph attention for referring expression comprehension. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 4643–4652. [doi: 10.1109/ICCV.2019.00474]

[17] Wang P, Wu Q, Cao JW, Shen CS, Gao LL, van den Hengel A. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 1960–1968. [doi: 10.1109/CVPR.2019.00206]

© 中国科学院软件研究所 http://www.jos.org.cn

[18] Yang SB, Li GB, Yu YZ. Relationship-embedded representation learning for grounding referring expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2765–2779. [doi: 10.1109/TPAMI.2020.2973983]

[19] Yang L, Xu Y, Yuan CF, Liu W, Li B, Hu WM. Improving visual grounding with visual-linguistic verification and iterative reasoning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 9489–9498. [doi: 10.1109/CVPR52688.2022.00928]

[20] Ye JB, Tian JF, Yan M, Yang XS, Wang XW, Zhang J, He L, Lin X. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 15481–15491. [doi: 10.1109/CVPR52688.2022.01506]

[21] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.

[22] Liao Y, Liu S, Li GB, Wang F, Chen YJ, Qian C, Li B. A real-time cross-modality correlation filtering method for referring expression comprehension. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 10877–10886. [doi: 10.1109/CVPR42600.2020.01089]

[23] Yang ZY, Chen TL, Wang LB, Luo JB. Improving one-stage visual grounding by recursive sub-query construction. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 387–404. [doi: 10.1007/978-3-030-58568-6_23]

[24] Huang BB, Lian DZ, Luo WX, Gao SH. Look before you leap: Learning landmark features for one-stage visual grounding. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 16883–16892. [doi: 10.1109/CVPR46437.2021.01661]

[25] Liao Y, Zhang AY, Chen ZY, Hui TR, Liu S. Progressive language-customized visual feature learning for one-stage visual grounding. IEEE Trans. on Image Processing, 2022, 31: 4266–4277. [doi: 10.1109/TIP.2022.3181516]

[26] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[27] Deng JJ, Yang ZY, Chen TL, Zhou WG, Li HQ. TransVG: End-to-end visual grounding with Transformers. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 1749–1759. [doi: 10.1109/ICCV48922.2021.00179]

[28] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2018. 4171–4186. [doi: 10.18653/v1/N19-1423]

[29] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[30] Tang KH, Niu YL, Huang JQ, Shi JX, Zhang HW. Unbiased scene graph generation from biased training. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3713–3722. [doi: 10.1109/CVPR42600.2020.00377]

[31] Zhang D, Zhang HW, Tang JH, Hua XS, Sun QR. Causal intervention for weakly-supervised semantic segmentation. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 56.

[32] Chen L, Yan X, Xiao J, Zhang HW, Pu SL, Zhuang YT. Counterfactual samples synthesizing for robust visual question answering. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10797–10806. [doi: 10.1109/CVPR42600.2020.01081]

[33] Yue ZQ, Zhang HW, Sun QR, Hua XS. Interventional few-shot learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 2734–2746.

[34] Wang T, Huang JQ, Zhang HW, Sun QR. Visual commonsense R-CNN. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10757–10767. [doi: 10.1109/CVPR42600.2020.01077]

[35] Wang T, Zhou C, Sun QR, Zhang HW. Causal attention for unbiased visual recognition. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 3071–3080. [doi: 10.1109/ICCV48922.2021.00308]

[36] Huang JQ, Qin Y, Qi JX, Sun QR, Zhang HW. Deconfounded visual grounding. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2022. 998–1006. [doi: 10.1609/aaai.v36i1.19983]

[37] Yang X, Zhang HW, Qi GJ, Cai JF. Causal attention for vision-language tasks. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9842–9852. [doi: 10.1109/CVPR46437.2021.00972]

[38] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[39] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]

[40] Liu Y, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized bert

© 中国科学院软件研究所 http://www.jos.org.cn

pretraining approach. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: ICLR, 2020. 1–15.

[41] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.

[42] Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RSZ, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the of the 32nd Int'l Conf. on Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.

[43] Hartigan JA, Wong MA. Algorithm AS 136: A K-means clustering algorithm. Applied Statistics, 1979, 28(1): 100. [doi: 10.2307/2346830]

[44] Kuhn HW. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 1955, 2(1–2): 83–97. [doi: 10.1002/nav.3800020109]

[45] Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 658–666. [doi: 10.1109/CVPR.2019.00075]

[46] Plummer BA, Wang LW, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2641–2649. [doi: 10.1109/ICCV.2015.303]

[47] Zhu DY, Chen J, Shen XQ, Li X, Elhoseiny M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: Proc. of the 12th Int'l Conf. on Learning Representations. Vienna: ICLR, 2024.

[48] Liu HT, Li CY, Wu QY, Lee YL. Visual instruction tuning. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 1516.

[49] Li JN, Li DX, Savarese S, Hoi S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 814.

[50] Bang Y, Cahyawijaya S, Lee N, Dai WL, Su D, Wilie B, Lovenia H, Ji ZW, Yu TZ, Chung W, Do QV, Xu Y, Fung P. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: Proc. of the 13th Int'l Joint Conf. on Natural Language Processing and the 3rd Conf. of the Asia-Pacific Chapter of the Association for Computational Linguistics (Vol. 1: Long Papers). Nusa Dua: Association for Computational Linguistics, 2023. 675–718. [doi: 10.18653/v1/2023.ijcnlp-main.45]

[51] Dong QX, Li L, Dai DM, Zheng C, Ma JY, Li R, XiaHM, Xu JJ, Wu ZY, Chang BB, Sun X, Li L, Sui ZF. A survey on in-context learning. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: Association for Computational Linguistics, 2022. 1107–1128. [doi: 10.18653/v1/2024.emnlp-main.64]

附中文参考文献:

[1] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. http://www.jos.org.cn/1000-9825/6125.htm [doi: 10.13328/j.cnki.jos.006125]

赵嘉宁(2000－), 男, 硕士生, CCF 学生会员, 主要研究领域为自然语言处理.

罗佳敏(1997－), 女, 博士生, CCF 学生会员, 主要研究领域为自然语言处理.

王晶晶(1990－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理.

周国栋(1967－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理.

© 中国科学院软件研究所 http://www.jos.org.cn