



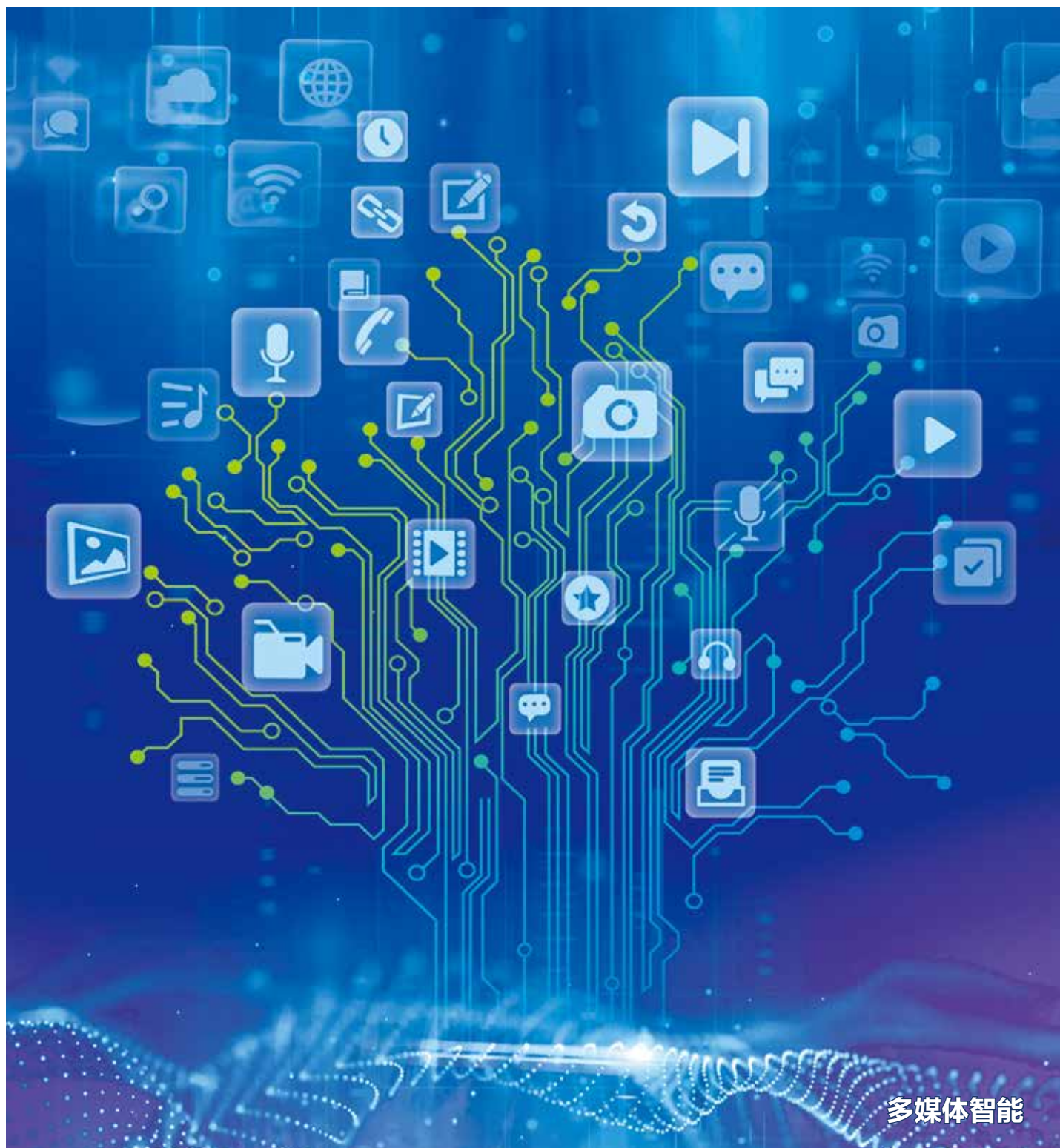
JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
09
VOL.27

ISSN1006-8961
CN11-3758/TB



多媒体智能



第27卷第9期（总第317期）
2022年9月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主 编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮 编 100190
电子信箱 jig@aircas.ac.cn
电 话 010-58887035
网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订 购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
(邮政信箱: 北京399信箱 邮编: 100048)
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

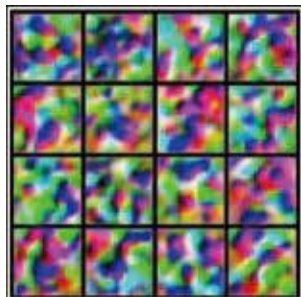
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



多媒体智能：当多媒体遇到人工智能(第2551页)



面向海洋的多模态智能计算：挑战、进展和展望(第2589页)



基于真实数据感知的模型功能窃取攻击(第2721页)

《中国图象图形学报》多媒体智能专刊简介

朱文武, 黄庆明, 黄华, 蒋树强, 彭宇新, 刘青山, 王井东, 纪荣嵘, 邓伟洪, 方玉明, 刘家瑛, 韩向娣 2549

学者观点

多媒体智能：当多媒体遇到人工智能

朱文武, 王鑫, 田永鸿, 高文 2551

视觉知识：跨媒体智能进化的新支点

杨易, 庄越挺, 潘云鹤 2574

面向海洋的多模态智能计算：挑战、进展和展望

聂婕, 左子杰, 黄磊, 王志刚, 孙正雅, 仲国强, 王鑫, 王玉成, 刘安安, 张弘, 董军宇, 魏志强 2589

综述

基于深度学习的人—物交互关系检测综述

廖越, 李智敏, 刘侃 2611

人类面部重演方法综述

刘锦, 陈鹏, 王茜, 付晓蒙, 戴娇, 韩冀中 2629

视觉语言多模态预训练综述

张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞 2652

Bayer阵列图像去马赛克算法综述

魏凌云, 孙帮勇 2683

多媒体智能安全

多特征决策融合的音频copy-move篡改检测与定位

张国富, 肖锐, 苏兆品, 廉晨思, 岳峰 2697

多级特征全局一致性的伪造人脸检测

杨少聪, 王健, 孙运莲, 唐金辉 2708

基于真实数据感知的模型功能窃取攻击

李延铭, 李长升, 余佳奇, 袁野, 王国仁 2721

目标智能检测

利用时空特征编码的单目标跟踪网络

王蒙蒙, 杨小倩, 刘勇 2733

结合时空一致性的FairMOT跟踪算法优化

彭嘉淇, 王涛, 陈柯安, 林巍峭 2749

多媒体分析与理解

融合知识表征的多模态Transformer场景文本视觉问答方法

余宙, 俞俊, 朱俊杰, 匡振中 2761

结合多层级解码器和动态融合机制的图像描述

姜文晖, 占锟, 程一波, 夏雪, 方玉明 2775

面向非受控场景的人脸图像正面化重建

辛经纬, 魏子凯, 王楠楠, 李洁, 高新波 2788

CONTENTS

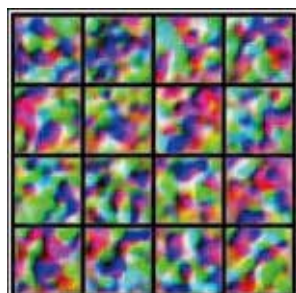
JOURNAL OF IMAGE AND GRAPHICS



Multimedia intelligence: the convergence of multimedia and artificial intelligence(P2551)



Marine oriented multimodal intelligent computing: challenges, progress and prospects(P2589)



Model functionality stealing attacks based on real data awareness(P2721)

Scholar View

Multimedia intelligence: the convergence of multimedia and artificial intelligence

Zhu Wenwu, Wang Xin, Tian Yonghong, Gao Wen 2551

The review of visual knowledge: a new pivot for cross-media intelligence evolution

Yang Yi, Zhuang Yueting, Pan Yunhe 2574

Marine oriented multimodal intelligent computing: challenges, progress and prospects

Nie Jie, Zuo Zijie, Huang Lei, Wang Zhigang, Sun Zhengya, Zhong Guoqiang, Wang Xin, Wang Yucheng, Liu An'an, Zhang Hong, Dong Junyu, Wei Zhiqiang 2589

Review

A review of deep learning based human-object interaction detection

Liao Yue, Li Zhimin, Liu Si 2611

Critical review of human face reenactment methods

Liu Jin, Chen Peng, Wang Xi, Fu Xiaomeng, Dai Jiao, Han Jizhong 2629

Comprehensive review of visual-language-oriented multimodal pre-training methods

Zhang Haoyu, Wang Tianbao, Li Mengze, Zhao Zhou, Pu Shiliang, Wu Fei 2652

The review of demosaicing methods for Bayer color filter array image

Wei Lingyun, Sun Bangyong 2683

Multimedia Intelligent Security

Multi-feature decision fused detection and localization method for copy-move forgery of digital audio clips

Zhang Guofu, Xiao Rui, Su Zhaopin, Lian Chensi, Yue Feng 2697

Multi-level features global consistency for human facial deepfake detection

Yang Shaocong, Wang Jian, Sun Yunlian, Tang Jinhui 2708

Model functionality stealing attacks based on real data awareness

Li Yanming, Li Changsheng, Yu Jiaqi, Yuan Ye, Wang Guoren 2721

Object Intelligent Detection

A spatio-temporal encoded network for single object tracking

Wang Mengmeng, Yang Xiaoqian, Liu Yong 2733

Spatio-temporal consistency based FairMOT tracking algorithm optimization

Peng Jiaqi, Wang Tao, Chen Kean, Lin Weiyao 2749

Multimedia Analysis and Understanding

Knowledge-representation-enhanced multimodal Transformer for scene text visual question answering

Yu Zhou, Yu Jun, Zhu Junjie, Kuang Zhenzhong 2761

The integrated mechanism of hierarchical decoders and dynamic fusion for image captioning

Jiang Wenhui, Zhan Kun, Cheng Yibo, Xia Xue, Fang Yuming 2775

Face frontalization for uncontrolled scenes

Xin Jingwei, Wei Zikai, Wang Nannan, Li Jie, Gao Xinbo 2788

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2022)09-2761-14

论文引用格式: Yu Z, Yu J, Zhu J J and Kuang Z Z. 2022. Knowledge-representation-enhanced multimodal Transformer for scene text visual question answering. Journal of Image and Graphics, 27(09): 2761-2774 (余宙, 俞俊, 朱俊杰, 匡振中. 2022. 融合知识表征的多模态 Transformer 场景文本视觉问答. 中国图象图形学报, 27(09): 2761-2774) [DOI: 10.11834/jig.211213]

融合知识表征的多模态 Transformer 场景文本视觉问答

余宙, 俞俊*, 朱俊杰, 匡振中

杭州电子科技大学计算机学院复杂系统建模与仿真教育部重点实验室, 杭州 310018

摘要: **目的** 现有视觉问答方法通常只关注图像中的视觉物体, 忽略了对图像中关键文本内容的理解, 从而限制了图像内容理解的深度和精度。鉴于图像中隐含的文本信息对理解图像的重要性, 学者提出了针对图像中场景文本理解的“场景文本视觉问答”任务以量化模型对场景文字的理解能力, 并构建相应的基准评测数据集 TextVQA (text visual question answering) 和 ST-VQA (scene text visual question answering)。本文聚焦场景文本视觉问答任务, 针对现有基于自注意力模型的方法存在过拟合风险导致的性能瓶颈问题, 提出一种融合知识表征的多模态 Transformer 的场景文本视觉问答方法, 有效提升了模型的稳健性和准确性。**方法** 对现有基线模型 M4C (multimodal multi-copy mesh) 进行改进, 针对视觉对象间的“空间关联”和文本单词间的“语义关联”这两种互补的先验知识进行建模, 并在此基础上设计了一种通用的知识表征增强注意力模块以实现两种关系的统一编码表达, 得到知识表征增强的 KR-M4C (knowledge-representation-enhanced M4C) 方法。**结果** 在 TextVQA 和 ST-VQA 两个场景文本视觉问答基准评测集上, 将本文 KR-M4C 方法与最新方法进行比较。本文方法在 TextVQA 数据集中, 相比于对比方法中最好的结果, 在不增加额外训练数据的情况下, 测试集准确率提升 2.4%, 在增加 ST-VQA 数据集作为训练数据的情况下, 测试集准确率提升 1.1%; 在 ST-VQA 数据集中, 相比于对比方法中最好的结果, 测试集的平均归一化 Levenshtein 相似度提升 5%。同时, 在 TextVQA 数据集中进行对比实验以验证两种先验知识的有效性, 结果表明提出的 KR-M4C 模型提高了预测答案的准确率。**结论** 本文提出的 KR-M4C 方法的性能在 TextVQA 和 ST-VQA 两个场景文本视觉问答基准评测集上均有显著提升, 获得了在该任务上的最好结果。

关键词: 场景文本视觉问答; 知识表征; 注意力机制; Transformer; 多模态融合

Knowledge-representation-enhanced multimodal Transformer for scene text visual question answering

Yu Zhou, Yu Jun*, Zhu Junjie, Kuang Zhenzhong

Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology,
Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: **Objective** Deep neural networks technology promotes the research and development of computer vision and natu-

收稿日期: 2022-01-05; 修回日期: 2022-06-01; 预印本日期: 2022-06-08

* 通信作者: 俞俊 yujun@hdu.edu.cn

基金项目: 国家重点研发计划资助 (2020YFB1406701); 国家自然科学基金项目 (62072147, 62125201); 浙江省自然科学基金项目 (LR22F020001, LY22F020028)

Supported by: National Key R&D Program of China (2020YFB1406701); National Natural Science Foundation of China (62072147, 62125201); Natural Science Foundation of Zhejiang Province, China (LR22F020001, LY22F020028)

ral language processing intensively. Multiple applications like human face recognition, optical character recognition (OCR), and machine translation have been widely used. Recent development enable the machine learning to deal with more complex multimodal learning tasks that involve vision and language modalities, e. g., visual captioning, image-text retrieval, referring expression comprehension, and visual question answering (VQA). Given an arbitrary image and a natural language question, the VQA task is focused on an image-content-oriented and question-guided understanding of the fine-grained semantics and the following complex reasoning answer. The VQA task tends to be as the generalization of the rest of multimodal learning tasks. Thus, an effective VQA algorithms is a key step toward artificial general intelligence (AGI). Recent VQA have realized human-level performance on the commonly used benchmarks of VQA. However, most existing VQA methods are focused on visual objects of images only, while neglecting the recognition of textual content in the image. In many real-world scenarios, the image text can transmit essential information for scene understanding and reasoning, such as the number of a traffic sign or the brand awareness of a product. The ignorance of textual information is constraint of the applicability of the VQA methods in practice, especially for visually-impaired users. Due to the importance of the developing textual information for image interpretation, most researches intend to incorporate textual content into VQA for a scene text VQA task organizing. Specifically, the questions involve the textual contents in the scene text VQA task. The learned VQA model is required to establish unified associations among the question, visual object and the scene text. The reasoning is followed to generate a correct answer. To address the scene text VQA task, a model of multimodal multi-copy mesh (M4C) is facilitated based on the transformer architecture. Multimodal heterogeneous features are as input, a multimodal transformer is used to capture the interactions between input features, and the answers are predicted in an iterative manner. Despite of the strengths of M4C, it still has the two weaknesses as following: 1) the relative spatial relationship cannot be illustrated well between paired objects although each visual object and OCR object encode its absolute spatial location. It is challenged to achieve accurate spatial reasoning for M4C model; 2) the predicted words of answering are selected from either a dynamic OCR vocabulary or a fixed answer vocabulary. The semantic relationships are not explicitly considered in M4C between the multi-sources words. At the iterative answer prediction stage, it is challenged to understand the potential semantic associations between multiple sources derived words. **Method** To resolve the weaknesses of M4C mentioned above, we improve the reference of M4C model by introducing two added knowledge like the spatial relationship and semantic relationship, and a knowledge-representation-enhanced M4C (KR-M4C) approach is demonstrated to integrate the two types of knowledge representations simultaneously. Additionally, the spatial relationship knowledge encodes the relative spatial positions between each paired object (including the visual objects and OCR objects) in terms of their bounding box coordinates. The semantic relationship knowledge encodes the semantic similarity between the text words and the predicted answer words in accordance with the similarity calculated from their GloVe word embeddings. The two types of knowledge representation are encoded as unified knowledge representations. To match the knowledge representations adequately, the multi-head attention (MHA) module of M4C is modified to be a KRMHA module. By stacking the KRMHA modules in depth, the KR-M4C model performs spatial and semantic reasoning to improve the model performance over the reference M4C model. **Result** The KR-M4C approach is verified that our extended experiments are conducted on two benchmark datasets of text VQA (TextVQA) and scene text VQA (ST-VQA) based on same experimental settings. The demonstrated results are shown as below: 1) excluded of extra training data, KR-M4C obtains an accuracy improvement of 2.4% over existing optimization on the test set of TextVQA; 2) KR-M4C achieves an average normalized levenshtein similarity (ANLS) score of 0.555 on the test set of ST-VQA, which is 5% higher than the result of SA-M4C. To verify the synergistic effect of two types of introduced knowledge further, comprehensive ablation studies are carried out on TextVQA, and the demonstrated results can support our hypothesis of those two types of knowledge are proactively and mutually to model performance. Finally, some visualized cases are provided to verify the effects of the two introduced knowledge representations. The spatial relationship knowledge improve the ability to localize key objects in the image, whilst the improved semantic relationship knowledge is perceived of the contextual words via the iterative answer decoding. **Conclusion** A novel KR-M4C method is introduced for the scene text VQA task. KR-M4C has its priority for the knowledge enhancement beyond the TextVQA and ST-VQA datasets.

Key words: scene text visual question answering; knowledge representation; attention mechanism; Transformer; multimodal fusion

0 引言

视觉问答(visual question answering, VQA) (Antol 等, 2015) 是计算机视觉和自然语言处理的交叉方向的典型任务, 也是近年来相关领域的研究热点。它以一幅图像和一个问题作为输入, 旨在设计模型对多模态的输入进行信息融合与推理, 最终以自然语言的形式输出问题的答案。图 1(a) 展示了视觉问答任务的一个示例。模型对图像和问题细致的理解后进行推理, 从而实现准确的答案预测, 因此它是一项极具挑战性的任务。近年来, 研究人员在视觉问答任务中取得了重大进展, 在一些常用的视觉问答基准测试集上 (Antol 等, 2015) 取得了接近人类水平的准确率, 但是这些方法大多忽视了对图像中“场景文本”这一重要信息的理解, 从而限制了其对场景理解的深度。另外, 视觉问答技术在现实生活中的一个典型应用场景是视障人群的辅助, 而对于这些特殊人群来说, 理解场景中的文字也是他们真正关心的痛点问题。

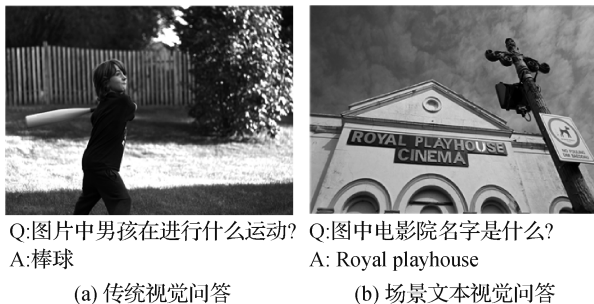


图1 传统视觉问答和场景文本视觉问答任务示例

Fig. 1 Examples of VQA and scene TextVQA

((a) VQA; (b) scene TextVQA)

Singh 等人 (2019) 和 Biten 等人 (2019) 提出将文本内容融入到视觉问答中, 形成面向场景文本视觉问答任务, 同时构建了 TextVQA (text visual question answering) (Singh 等, 2019) 和 ST-VQA (scene text visual question answering) (Biten 等, 2019) 两个基准数据集。图 1(b) 展示了场景文本视觉问答任务的一个示例, 该任务问题涉及图像中相关的场景文本, 需要模型建立问题、视觉对象和场景文本之间的统一关联后开展推理以生成正确的答案。为了理解图像中的场景文本, 场景文本视觉问答模型通常需要引入一个光学字符识别 (optical character recog-

nition, OCR) 系统来检测并识别图像中的文本对象。基于抽取到的 OCR 对象, 一些方法相继提出 (Singh 等, 2019; Hu 等, 2020; Kant 等, 2020)。LoRRA (look read reason and answer) (Singh 等, 2019) 方法在视觉问答模型的基础上, 扩展了一个用于场景文本编码的 OCR 注意分支。M4C (multimodal multi-copy mesh) (Hu 等, 2020) 方法通过多模态 Transformer 融合所有输入特征模态内和模态间的信息, 并采用迭代解码生成答案。但是 M4C 中 Transformer 的自注意力层是完全连接的, 将注意力分散到整体上下文中, 而忽略了围绕特定对象或文本的局部上下文的重要性。

在场景文本视觉问答任务中, 部分问题涉及推理对象间的相对空间关系。例如, 图像右侧的标识牌上写了什么内容? 或图像左边球员的球衣上写了什么数字? 针对这类问题, SA-M4C (spatially aware M4C) (Kant 等, 2020) 方法通过引入 12 种预先定义的空间关系 (Yao 等, 2018) 对视觉对象和 OCR 对象构建联系, 获得了增强的相对空间关系知识, 并将其融合到 Transformer 每个注意力层中, 改进并提升了 M4C 方法的性能。但是人工构建的空间关系的空间量化策略不够精准, 对于空间关联紧密的目标表达不够精准。

本文提出了一种融合知识表征的多模态 Transformer 的场景文本视觉问答方法 KR-M4C (knowledge-representation-enhanced M4C), 通过将“空间关联”和“语义关联”两种互补的先验知识进行统一建模后融入 M4C 模型框架, 提升模型对复杂场景的理解能力。空间关联知识对视觉对象和 OCR 对象间的相对空间位置进行编码表征, 有效对两两对象间细粒度的空间关系进行精准刻画。语义关联知识对 OCR 对象对应的单词和预测答案单词之间的语义相似性, 对存在上下文语义关联的单词进行编码表征, 提升答案生成过程模型的准确性和可靠性。为了评估提出的 KR-M4C 方法的有效性, 分别在 TextVQA 和 ST-VQA 两个公开数据集上进行实验。实验结果表明, 相较于目前最好的方法, KR-M4C 在两个数据集上均取得显著的性能提升。本文主要贡献如下: 1) 提出一种基于知识表征增强的多模态 Transformer 场景文本视觉问答模型 KR-M4C, 通过引入增强的知识, 获得了更丰富的信息表示; 2) 建模了视觉对象间“空间关联”和单词间的“语义关

联”这两种互补的先验知识,更准确地引导模型定位关键物体对象和文本对象;3)在 TextVQA 数据集和 ST-VQA 数据集上进行了充分的对比实验和消融实验。实验结果表明, KR-M4C 与现有最好方法相比具有更出色的表现。

1 相关工作

多模态学习旨在设计模型对来自不同模态的信号(如视觉、听觉和语言等)进行信息关联,并在此基础上学习统一的语义表达。得益于深度学习的快速发展,多模态学习逐渐成为计算机视觉和自然语言处理领域的研究热点,提出了一系列重要的多模态学习任务。如视觉问答(Antol 等,2015;Kim 等,2018;Yu 等,2019)、视觉定位(Yu 等,2017a)、图文检索(Karpathy 和 Li,2015;Lee 等,2018)和图像描述(Anderson 等,2016;Veit 等,2016)等。在这些任务中,视觉问答是一个典型且具有挑战性的任务。

VQA 是多模态领域近年来的研究热点。VQA 任务的核心在于如何进行图像和问题的多模态信息融合。Zhou 等人(2015)和 Antol 等人(2015)使用的是特征拼接或对应元素相加的线性融合方法。而后,得益于双线性模型在细粒度识别领域取得的良好效果,Fukui 等人(2016)、Kim 等人(2017)、Ben-Younes 等人(2017)和 Yu 等人(2017b)设计了不同的近似双线性池化模型用于多模态特征的细致语义融合。随着对注意力机制研究的深入,提出了深度共同注意力模型用于多模态融合和注意力学习。Yu 等人(2019)将模块化的思想引入视觉问答中,设计两种注意力单元并进行模块化组合,构建深度模块化协同注意力网络,建模多模态数据间细粒度的交互关联。随着多模态预训练研究的兴起,研究人员不再聚焦于单一的视觉问答模型设计,而是将研究重心聚焦如何基于 Transformer 这种通用架构设计合适的预训练策略,在大规模数据集上进行预训练,并在得到的模型权重基础上利用 VQA 数据进行模型权重微调。Lu 等人(2019)、Tan 和 Bansal(2019)、Chen 等人(2020)和 Cui 等人(2021)提出了各种多模态预训练框架,并逐步刷新视觉问答任务基准评测集上的最好成绩。

场景文本视觉问答任务是一种特殊的视觉问答任务。相比一般视觉问答任务,场景文本视觉问答

更侧重图像中的文本信息,不仅它的问题主要围绕图像中的文本信息,而且它的回答也需要使用图像中的文本信息。这要求模型对问题单词、图像的物体对象和图像的文本对象构建联系,并经过推理生成正确答案。由于该任务具有重要的研究和应用价值,研究人员提出了一系列的解决方案。LoRRA(Singh 等,2019)扩展了面向传统视觉问答的 Pythia(Jiang 等,2018)方法,使用现有的 OCR 系统检测图像中的文本对象,额外增加了一个用于场景文本编码的 OCR 注意分支,通过对固定词汇表中的单词和 OCR 识别得到的单词进行推理,选择其中概率最大的单词作为答案。由于 LoRRA 没有构建丰富的 OCR 对象特征,因此无法充分理解 OCR 对象蕴含的信息。此外,其采用的浅层注意力融合模型无法进行深度推理。为解决该问题,Hu 等人(2020)提出了 M4C 模型,该模型是该任务上的一个强有力的基线方法,它通过一个多模态 Transformer(Antol 等,2015)结构,将不同模态的特征嵌入到一个共同的语义空间中进行融合。在这个语义空间中,模态间和模态内的关联由自注意力模型自动学习得到。M4C 将场景文本视觉问答作为一个序列生成任务,结合一个动态指针网络模块实现精准的答案生成。然而,M4C 没有考虑到图像中对象之间的相对空间关系,因此在涉及相对空间关系的问题上表现不理想。为解决 M4C 模型存在的问题,SA-M4C(Kant 等,2020)在 M4C 的基础上,通过 12 种空间关系(Yao 等,2018)对物体对象和 OCR 文本对象构建联系,获得了增强的知识,进一步提升了准确率,但是预定义的空间关系建模方法对空间关系的表达粒度不够细致,对相近位置对象难以区分。

2 融合知识表征多模态 Transformer

针对场景文本视觉问答任务,本文在 M4C 模型(Hu 等,2020)基础上进行改进,通过对“空间关系”和“语义关系”两种互补先验知识进行联合建模,构建知识表征增强的 M4C 模型。

2.1 M4C 模型概述

M4C 模型以问题和图像两种模态的数据作为输入,在进行多模态数据统一表征后,输入一个多层 Transformer 网络对多模态信息深度融合,最后输入一个动态指针网络进行迭代式答案预测。

2.1.1 多模态数据统一表征

如图2所示,M4C对问题和答案中的单词分别进行词向量表征,对图像分别提取视觉对象区域和OCR对象区域,并针对提取区域进行视觉对象表征和OCR对象表征。

给定由至多 K 个单词组成的问题,使用经过预训练的BERT-base(bidirectional encoder representation from transformers-base)模型(Devlin等,2019)将每个单词映射到 d 维特征向量,最后将每个单词的表征拼接后得到问题特征 $\mathbf{X}^{\text{ques}} \in \mathbf{R}^{K \times d}$ 。

给定一幅图像,使用经过预训练的Faster R-CNN(region convolutional neural network)模型(Ren等,2015)从图像中提取至多 M 个视觉对象区域。对第 m 个视觉对象(其中 $m \in \{1, \dots, M\}$)提取视觉特征 $\mathbf{x}_m^{\text{vis}}$ 和坐标特征 $\mathbf{x}_m^{\text{box}}$ 。然后使用线性映射将上述两种特征统一映射到 d 维空间,得到视觉对象表征 $\mathbf{x}_m^{\text{obj}}$ 。最后将 M 个视觉对象表征对应元素后,得到 $\mathbf{X}^{\text{obj}} \in \mathbf{R}^{M \times d}$ 。

针对图像中的文本,首先使用外部OCR系统从

图像中提取至多 N 个文本对象区域并得到相应的单词表示。对第 n 个文本对象(其中 $n \in \{1, \dots, N\}$)抽取如下4种特征:基于Fast-Text方法(Bojanowski等,2017)的词向量特征、基于PHOC(pyramidal histogram of characters)方法(Almazán等,2014)的字符特征、基于Faster R-CNN的视觉特征和文本区域的包围框坐标特征。然后使用线性映射将上述4种特征统一映射到 d 维空间,得到每个文本对象表征 $\mathbf{x}_n^{\text{ocr}}$ 。最后将 N 个文本对象表征拼接后得到 $\mathbf{X}^{\text{ocr}} \in \mathbf{R}^{N \times d}$ 。

对于由至多 T 个单词组成的答案序列,将每个单词映射到 d 维向量 $\mathbf{x}_t^{\text{ans}}$ (其中 $t \in \{1, \dots, T\}$),最后将 T 个单词表征进行拼接后得到 $\mathbf{X}^{\text{ans}} \in \mathbf{R}^{T \times d}$ 。在答案词汇表中,本文额外增加两个特殊的单词[start]和[end],对应答案生成过程中的起始符和终止符。

将上述4组多模态特征 $\mathbf{X}^{\text{ques}}$ 、 $\mathbf{X}^{\text{obj}}$ 、 $\mathbf{X}^{\text{ocr}}$ 和 $\mathbf{X}^{\text{ans}}$ 通过拼接得到融合特征 $\mathbf{Z} \in \mathbf{R}^{(K+M+N+T) \times d}$,然后将其输入层的Transformer编码器模型(Vaswani等,2017)进行特征的深度融合。

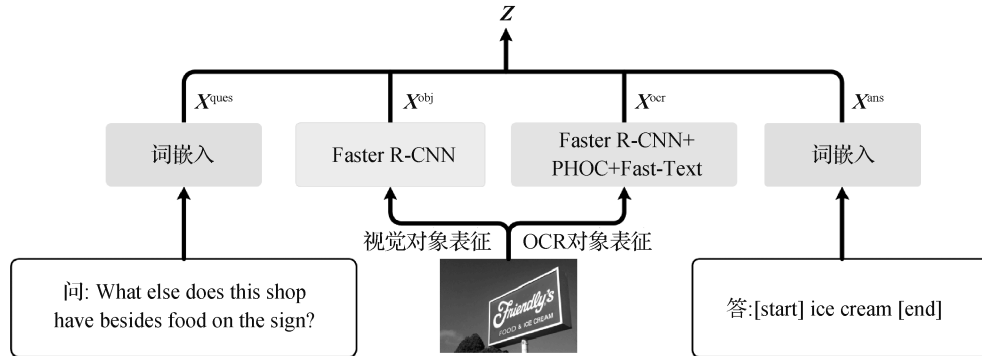


图2 M4C多模态表征流程示意图

Fig.2 Flowchart of the multimodal representations of M4C

2.1.2 自回归答案预测

由于答案可能包含OCR单词与词汇表单词合并形成的短语,M4C引入结合动态指针网络的自回归答案预测模块,将在OCR单词和预先构建的答案词汇表中进行的自适应选择作为 t 时刻的预测输出,并将本次预测结果作为下一次解码的输入,迭代解码,直至输出终止符。

令 z_t^{ans} 表示第 t 时刻答案单词经由Transformer编码器的输出特征,预设的答案词汇表由 V 个单词构成。 z_t^{ans} 经由一个线性层映射,得到 V 维的词汇表单词得分向量 $\mathbf{y}_t^{\text{voc}} \in \mathbf{R}^V$ 。此外,令 $\{z_1^{\text{ocr}}, \dots, z_N^{\text{ocr}}\}$ 表

示 N 个OCR单词经过Transformer得到的特征,动态指针网络(dynamic pointer network,DPN)模块旨在计算 z_t^{ans} 和任意OCR单词之间的关联性得分,计算为

$$\mathbf{y}_{t,n}^{\text{ocr}} = \text{line}_d(z_t^{\text{ans}})^T \text{line}_d(z_n^{\text{ocr}}) \quad (1)$$

式中, $\text{line}(\cdot)$ 表示输出 d 维的线性映射层, $\mathbf{y}_{t,n}^{\text{ocr}}$ 为第 t 时刻答案单词和第 n 个OCR单词的得分。通过对所有OCR单词进行计算,得到OCR得分向量 $\mathbf{y}_t^{\text{ocr}} \in \mathbf{R}^N$ 。

最后,将上述两个向量进行拼接得到预测向量,通过计算预测答案与正确答案的BCE(binary cross-

entropy) 累计损失,实现对整个 M4C 模型的端到端优化。

2.2 KR-M4C 模型的整体结构

得益于多层 Transformer 模型强大的建模表达能力, M4C 模型可以学习得到不同模态特征之间的细粒度语义关联。但其存在两个弱点: 1) 尽管每个视觉对象和 OCR 对象中包含了其对应空间位置信息, 然而这种空间位置特征与其他类型特征进行特征融合失去了明确的空间坐标含义, 使得 M4C 模型难以准确理解对象间的空间关系; 2) 在进行答案预测时输出的单词需要从 OCR 对象单词和词汇表单词中进行选择, 而这些不同来源的单词之间的语义关联在 M4C 中并未显式地建模, 在进行答案预测过程中难以准确理解多来源单词之间的语义关联。

为了解决上述问题, 本文在图像中的视觉对象和 OCR 对象之间提取空间关联知识表征, 在 OCR 对象提取得到的单词和预测单词之间提取语义关联知识表征, 并针对这两种关联知识表征, 在 M4C 架构基础上提出一种知识表征增强的改进方法 KR-M4C 对知识进行编码表达。KR-M4C 整体框架如图 3 所示, 视觉对象与 OCR 对象间的“空间关联”、OCR 单词与预测答案单词间的“语义关联”有助于提升场景文本视觉问答的准确性。

1) 空间关联知识表征。场景文本视觉问答需要模型理解图像中视觉对象与 OCR 对象之间的空间关系, 并在此基础上进行推理。如图 4 所示, 给定问题 What's the licence number of the car? 模型首先要检测到图像中的车牌对应的视觉对象, 然后推断出号码与车牌之间的关系, 判断只有车牌对象内的

数字才是需要回答的车牌号。

给定图像抽取得到的 M 个视觉对象特征 X^{obj} 和 N 个 OCR 对象特征 X^{ocr} , 将其看做 $M + N$ 个元素构成的对象集合, 并将任意两两元素之间的相对空间关联信息进行编码, 得到空间关联知识表征。具体而言, 给定任意两个对象的对应包围框 (bounding box) 信息, 现有方法 SA-M4C (Kant 等, 2020) 提出一种基于空间关系图的空间位置编码方式, 通过预先定义的 12 种常见的空间关系 (如上下左右、里外、包含和重叠等) 表达任意两个对象之间的空间相对位置。但这种方法对空间关系的表达粒度不够细致, 对相近位置对象难以区分。为此, 本文采用 Hu 等人 (2018) 提出的基于坐标位置的相对位置编码方法解决该问题。令第 i 个对象的包围框表示为 (x_i, y_i, w_i, h_i) , 其中 x_i, y_i 分别表示包围框的中心点横坐标和纵坐标, w_i, h_i 分别表示包围框宽度和高度。第 i 和第 j 个对象之间的空间关系特征 $x_{i,j}^{sp} \in \mathbf{R}^4$, 计算为

$$x_{i,j}^{sp} = \left[\log \frac{|x_i - x_j|}{w_i}, \log \frac{|y_i - y_j|}{h_i}, \log \frac{w_j}{w_i}, \log \frac{h_j}{h_i} \right] \quad (2)$$

对于 $M + N$ 个对象, 使用上述方式得到相应的空间位置表征 $X^{sp} \in \mathbf{R}^{(M+N) \times (M+N) \times 4}$ 。进一步, 将 X^{sp} 使用两层全连接网络, 得到最终的空间关系知识表征 $R^{sp} \in \mathbf{R}^{(M+N) \times (M+N)}$ 。具体为

$$R^{sp} = \text{relu} \circ \text{line}_1 \circ \text{relu} \circ \text{line}_d (X^{sp}) \quad (3)$$

式中, relu 表示 ReLU 层, $\circ$ 表示两个层之间的顺序连接。

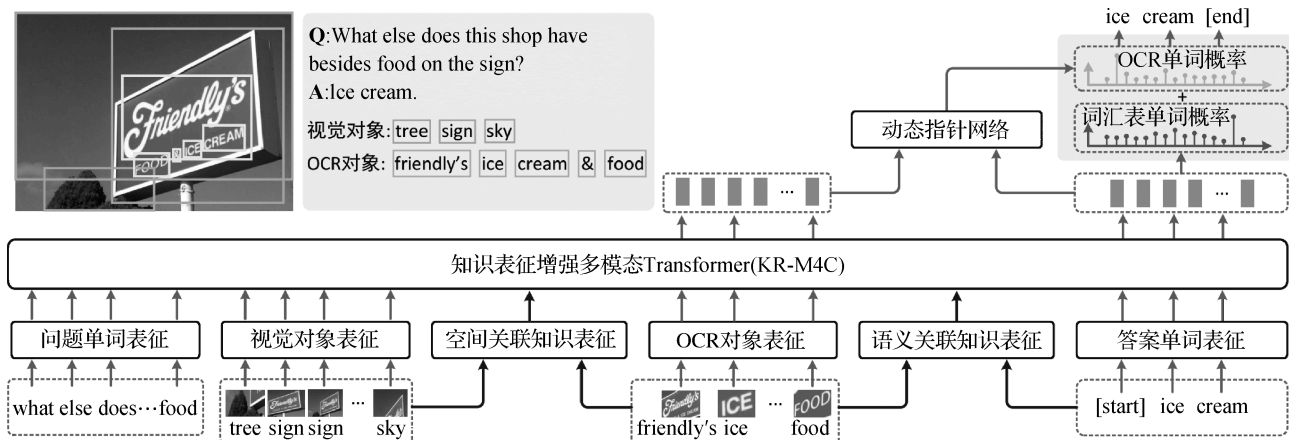


图3 KR-M4C 整体框架图

Fig. 3 An overview architecture of KR-M4C

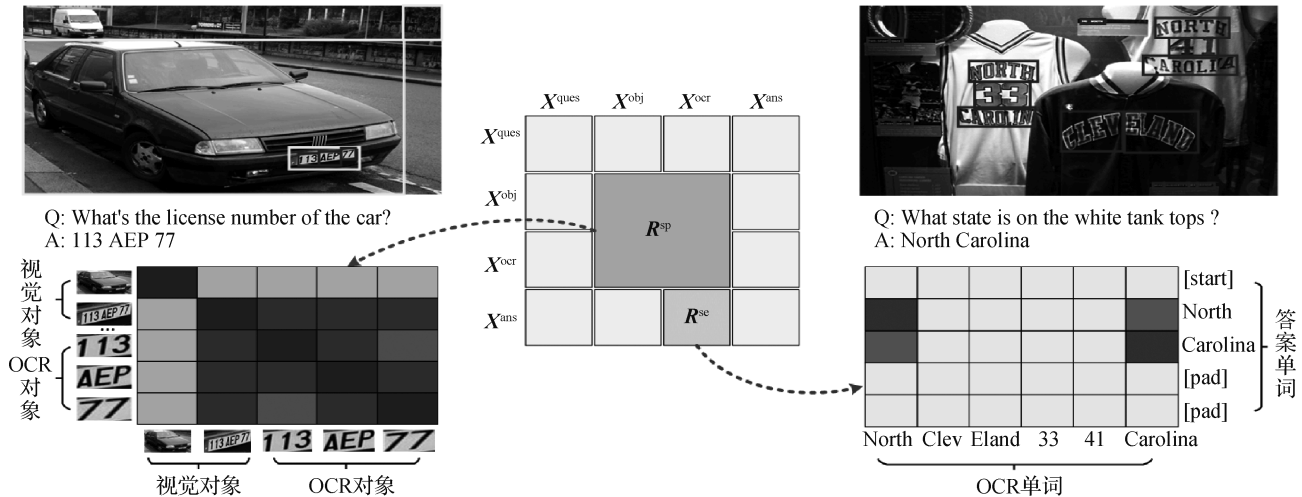


图4 空间关联知识表征与语义关联知识建模示意图

Fig. 4 Schematic diagram of spatial relationship knowledge representation and semantic relationship knowledge modeling

2) 语义关联知识表征。部分问题也可能会涉及空间上不相邻的多个单词组合,如图4所示的North和Carolina。如何去挖掘这些单词之间隐含的语义关系,这直接影响最后预测的精度。此外,由于答案中单词来源由OCR单词和固定词汇表两方面组成,并通过两个独立的分类器预测得到,因此需要模型学习理解两种不同来源单词之间的语义关联。

为解决上述问题,本文在OCR对象对应的单词与预测的答案输出单词之间构建语义关联知识,使用预训练的词向量表征之间的相似度计算单词之间的相对语义关系。具体而言,给定第 i 个OCR对象对应的单词和预测答案的第 j 个单词,它们对应的GloVe(global vectors for word representation)词向量(Pennington等,2014)表征为 e_i 和 e_j 。使用两个词向量间的余弦相似度(cossim)作为对应单词相对语义关系,即

$$x_{i,j}^{se} = f_{\text{cossim}}(e_i, e_j) \quad (4)$$

给定 N 个OCR对象和长度为 T 的答案,通过对所有答案单词和OCR单词进行计算,得到语义关联知识表征 $R^{se} \in \mathbf{R}^{T \times N}$ 。对于长度不足 T 的答案,使用一个特殊单词[pad]表示。[start]、[end]和[pad]这3个特殊单词与OCR单词之间的相似度不进行计算,并将其设置为0。

3) 知识表征增强自注意力模块。在原生M4C模型中,多模态融合得到的特征 Z 输入由 L 个自注意力模块堆叠而成的Transformer编码器。每个模块由一个多头注意力(multi-head attention, MHA)模

块和一个两层感知机构成的前向神经网络(feed-forward networks, FFN)模块堆叠而成。MHA利用 h 个并行的自注意力函数的输出特征拼接,形成多个注意力函数融合的多样性特征表达。其中,第 j 个自注意力函数定义为

$$\text{head}_j(Z) = SA(\text{line}_{d_h}(Z), \text{line}_{d_h}(Z), \text{line}_{d_h}(Z)) \quad (5)$$

$$SA(Q, K, V) = \text{softmax}(QK^T / \sqrt{d})V \quad (6)$$

式中, $d_h = d/h$ 为每个注意力函数的输出特征维度, softmax 表示 softmax 激活函数。

受Hu等人(2018)提出的空间关系建模方法的启发,本文将其应用在KR-M4C模型中,将式(5)中的自注意力函数改写为知识先验引导的自注意力函数。具体地,将 R^{sp} 和 R^{se} 合并后得到一个统一的知识表征矩阵 R 。为了使 R^{sp} 和 R^{se} 的数值范围相近,需要对 R^{sp} 进行变换,令 $R^{sp} = \log(R^{sp} + \varepsilon)$,其中 $\varepsilon = 10^{-6}$ 为预设的常数项,防止计算过程中的数值下溢。提出的知识表征增强自注意力函数KRSA(knowledge-representation-enhanced self-attention)为

$$\text{KRSA}(Q, K, V, R) = \text{softmax}(QK^T / \sqrt{d} + R)V \quad (7)$$

式中, Q 、 K 、 V 均对应特征个数为 $P = K + M + N + T$ 的多模态拼接后的特征 Z 。为了实现式(7)中的知识融合,得到知识表征 $R \in \mathbf{R}^{P \times P}$,将 R^{sp} 和 R^{se} 按照图4所示的方式进行排布, R 中其余位置使用0进行填充。需要指出的是, KRSA中 R^{sp} 和 R^{se} 的使用非常灵活,既可以支持两者同时输入,也可以

支持任意一种知识输入。此外,当 R 为全 0 元素填充时, KRSA 退化为标准的 SA (self-attention) 模块, 对应标准的 M4C 模型。通过将 KRSA 与式(5)中多头注意力机制结合, 得到知识表征增强的多头注意力模块 KRMHA (knowledge-representation-enhanced multi-head attention), 具体为

$$KRMHA(Z) = \text{line}_d([\text{head}_1(Z), \dots, \text{head}_h(Z)]) \quad (8)$$

4) KR-M4C 主干网络。通过将原生 M4C 模型中的 MHA 模块替换为知识表征增强的 KRMHA 模块, KR-M4C 的主干模型为 L 层的深度结构。具体为

$$\hat{Z}^l = \text{norm}(KRMHA(Z^{l-1}) + Z^{l-1}), l = 1, \dots, L \quad (9)$$

$$Z^l = \text{norm}(FFN(\hat{Z}^l) + \hat{Z}^l), l = 1, \dots, L \quad (10)$$

式中, $Z^0 = Z$, $\text{norm}()$ 表示层标准化处理, $FFN()$ 表示 2 层感知机构成的前馈神经网络。输出特征 Z^L 接入 M4C 中的自回归答案预测模块实现迭代式答案预测。

3 实验

为验证本文 KR-M4C 的有效性, 在场景文本视觉问答任务的 TextVQA 数据集和 ST-VQA 数据集上进行验证, 并与现有的场景文本视觉问答模型的结果进行对比, 对模型生成的样例进行分析。

3.1 实验设置

3.1.1 数据集

1) TextVQA 数据集 (Singh 等, 2019) 收集了来自 Open-Images v3 (Kuznetsova 等, 2020) 数据集的 28 408 幅自然场景图像, 其中训练集 21 953 幅, 验证集 3 166 幅, 测试集 3 289 幅。对各幅图像提出了 1~2 个问题, 共 45 336 个问题, 其中训练集包含 34 602 个问题, 验证集包含 5 000 个问题, 测试集包含 5 734 个问题。各问题有 10 个人工标注的答案, 通过 10 个答案的加权投票得分计算准确率。本文遵循先前的工作 (Hu 等, 2020), 在训练集的答案中收集前 5 000 个高频的单词构成固定词汇表。

2) ST-VQA 数据集 (Biten 等, 2019) 包含 MS COCO (Microsoft common objects in context) (Veit 等, 2016)、VizWiz (Gurari 等, 2018)、ICDAR (international

conference on document analysis and recognition) (Karatzas 等, 2015)、ImageNet (Deng 等, 2009)、Visual Genome (Krishna 等, 2017) 和 IIIT-STR (IIIT scene text retrieval) (Mishra 等, 2013) 这 6 个数据集共 21 892 幅自然场景图像, 其中训练集和验证集包含 18 921 幅图像, 测试集包含 2 971 幅图像。针对各幅图像, 设计了 1~3 个问题, 各问题提供 1~2 个人工标注的答案。该数据集涉及 3 项任务, 本文遵循先前的工作 (Hu 等, 2020), 在任务 3 上验证模型的效果, 原因是任务 3 与 TextVQA 数据集相似, 官方不提供候选单词, 需要通过外部 OCR 系统或固定词汇表得到候选单词。与 TextVQA 数据集不同, ST-VQA 数据集推荐采用 ANLS (average normalized levenshtein similarity) 得分 (Biten 等, 2019) 作为官方评价指标, 实现更为精准的答案一致性评估。在此基础上, 实验时采用 TextVQA 数据集上使用的准确率指标作为补充。

3.1.2 实现细节

实验环境为装载了 Nvidia Titan Xp GPU 的工作站。使用 PyTorch 框架 (Paszke 等, 2019) 实现 KR-M4C 模型。遵循先前的工作 (Kant 等, 2020), 模型设计和训练的超参数如表 1 所示。

表 1 KR-M4C 模型超参数

Table 1 Hyperparameter choices for KR-M4C

超参数	值
问题词最大数量 K	20
物体对象最大数量 M	100
OCR 对象最大数量 N	50
最长答案长度 T	12
统一表征维度 d	768
注意力头个数 h	12
词汇表大小 V	5 000
优化器	Adam
批次大小	128
基础学习率	10^{-4}
学习率衰减因子	0.1
学习率衰减轮数	14 000/19 000
迭代次数	24 000

3.2 消融实验

为了验证不同模型架构对 KR-M4C 的效果,在 TextVQA 数据集上进行消融实验。为公平对比,消融实验中的所有模型使用 Microsoft OCR 提取图像中的文本信息,并使用主干网络为 ResNeXt-152 的 Faster R-CNN 模型进行视觉特征表达 (Xie 等, 2017)。实验结果如表 2 所示。

表 2 TextVQA 数据集上的消融实验
Table 2 Ablation experiments on the TextVQA dataset

行号	方法	模型层数 (架构)	验证集 准确率/%
1	M4C (Hu 等, 2020)	6(N)	45.75
2	SA-M4C (Kant 等, 2020)	6(S)	46.93
3	KR-M4C	6(K)	48.42
4	KR-M4C 去除 R^{sp}	6(K)	46.90
5	KR-M4C 去除 R^{sc}	6(K)	47.06
6	SA-M4C (Kant 等, 2020) *	2(N) → 4(S)	46.70
7	KR-M4C	2(N) → 4(K)	47.91
8	KR-M4C	3(N) → 3(K)	47.04
9	KR-M4C	4(N) → 2(K)	47.18
10	KR-M4C	4(K)	46.10
11	KR-M4C	8(K)	48.34
12	KR-M4C	10(K)	48.31

注:加粗字体表示最优结果,括号中的 N 表示标准的注意力结构, S 表示 SA-M4C 中的空间感知注意力结构, K 表示本文提出的知识表征增强的注意力结构。* 表示基于原开源代码复现。

1) 不同类型知识的增强效果。以 M4C 模型架构为对照,验证不同类型知识引入后的增强效果。首先对比表 2 第 1、3 行结果,发现在模型层数相同(均为 6 层)情况下,本文提出的 KR-M4C 方法显著优于 M4C 方法 (Hu 等, 2020),证明了引入知识表征增强的有效性。其次对比第 1、3、4、5 行结果,发现仅建模空间关联知识 R^{sp} 或语义关联知识 R^{sc} 的方法均带来一定程度的性能下降,但相较 M4C 方法仍有一定提升,证明引入的两种知识均对模型具有一定程度的贡献,且两种知识具有良好的互补特性。最后对比第 2、5 行结果,尽管 SA-M4C (Kant 等, 2020) 和只包含空间关联 R^{sp} 的 KR-M4C 使用了相近的知识表征增强策略,但 SA-M4C 的空间关联建模粒度比 KR-M4C 更粗,因此其准确率稍逊于 KR-

M4C 方法。

2) 不同模块组合的影响。在 SA-M4C (Kant 等, 2020) 中,将不同类型的注意力模块进行组合,如 2 个 M4C 中标准的自注意力(N)和 4 个 SA-M4C 中的空间感知注意力模块(S),可以提升模型的表达能力。因此,在表 2 第 3、8—11 行的结果中,保持模型层数 $L = 6$ 不变,探索不同类型的注意力模块即 M4C 中的标准注意力模块(N)和 KR-M4C 中的知识表征增强注意力模块(K)的深度组合方式对结果的影响。引入知识的 6(K)架构的每一层都取得了最好效果,相比第 7—9 行中 $a(N) \rightarrow (6-a)(K)$ 架构准确率至少提升 0.5%。但该结果与 SA-M4C 论文中的结果并不一致。原因或是 SA-M4C 的空间信息编码策略约束过强,在前几层中会削弱对问题语义的理解。相比之下, KR-M4C 在每一层知识融合时引入了一组独立的可学习的参数,使模型可以自适应地学习在不同层中知识融合的程度,从而获得更好的融合效果。

3) 不同堆叠深度的影响。保持 $L(K)$ 的架构设计,探索不同深度 L 下模型的表现。从表 2 第 3、10、11 行结果可以看出,随着深度 L 从 4 增加到 10, KR-M4C 模型性能先上升然后逐渐下降。在 $L = 6$ 时取得最优结果。造成该现象的原因是过深的模型会导致模型优化困难,限制了模型的表达能力。该问题或许可以通过引入更多的训练数据得到缓解。

3.3 与现有方法的对比实验

基于消融实验的结果,使用 6(K) 的最优架构在 TextVQA 数据集上与目前最好方法对比,包括 LoRRA (Singh 等, 2019)、MM-GNN (multi-modal graph neural network) (Gao 等, 2020)、M4C (Hu 等, 2020)、SMA (structured multimodal attentions) (Gao 等, 2021)、CRN (cascade reasoning network) (Liu 等, 2020)、LaAPNet (localization-aware answer prediction network) (Han 等, 2020) 和 SA-M4C (Kant 等, 2020),结果如表 3 所示。额外训练数据表示除 TextVQA 数据集以外的数据,如 ST-VQA 数据集。

表 3 展示了不同条件下(如特征提取主干网络、OCR 系统和外部训练数据)的公平对比结果。特征提取主干网络包括基于 Faster R-CNN 的特征提取器与 ResNet 和 ResNeXt 两种主干网络的组合。

表 3 TextVQA 数据集上与现有最好方法的对比结果

Table 3 Comparative results with existing state-of-the-art methods on TextVQA dataset

行号	方法	特征提取主干网络	OCR 系统	额外训练数据	验证集准确率/%	测试集准确率/%
1	LoRRA(Singh 等,2019)	ResNet	Rosetta-ml	—	26.56	27.63
2	MM-GNN(Gao 等,2020)	ResNet	Rosetta-ml	—	31.44	31.10
3	M4C(Hu 等,2020)	ResNet	Rosetta-en	—	39.40	39.01
4	SMA(Gao 等,2021)	ResNet	Rosetta-en	—	40.05	40.66
5	CRN(Liu 等,2020)	ResNet	Rosetta-en	—	40.39	40.96
6	LaAPNet(Han 等,2020)	ResNet	Rosetta-en	—	40.68	40.54
7	KR-M4C(本文)	ResNet	Rosetta-en	—	41.78	42.99
8	M4C(Hu 等,2020)	ResNet	Rosetta-en	ST-VQA	40.55	40.46
9	LaAPNet(Han 等,2020)	ResNet	Rosetta-en	ST-VQA	41.02	41.41
10	KR-M4C(本文)	ResNet	Rosetta-en	ST-VQA	42.78	43.51
11	SA-M4C(Kant 等,2020)	ResNeXt	Google OCR	—	43.90	—
12	SA-M4C(Kant 等,2020)*	ResNeXt	Microsoft OCR	—	46.70	—
13	KR-M4C(本文)	ResNeXt	Microsoft OCR	—	48.42	—
14	SA-M4C(Kant 等,2020)*	ResNeXt	Microsoft OCR	ST-VQA	47.99	48.52
15	KR-M4C(本文)	ResNeXt	Microsoft OCR	ST-VQA	49.27	49.63

注:加粗字体为各不同对比条件的最优结果,“—”表示无实验数据,*表示基于原开源代码复现。

OCR 系统包括各方法采用的 OCR 系统。大部分方法沿用了 M4C 中的策略,使用 Facebook 提供的 Rosetta-ml 和 Rosetta-en 系统(Borisyuk 等,2018),SA-M4C 采用了识别性能更好的 Google OCR 系统。本文使用了总体效果更好的 Microsoft OCR 系统。从实验结果中可以得到如下结论:1)在使用 Rosetta-en OCR 系统、Faster R-CNN 主干网络使用 ResNet-101 的条件下,KR-M4C 在验证集和测试集上准确率为 41.78% 和 42.99%(第 7 行),相比最好结果提升 1.1% 和 2.5%;2)增加 ST-VQA 数据集作为额外训练数据,KR-M4C 在验证集、测试集上准确率分别为 42.78% 和 43.51%(第 10 行),均为同等条件下的最好结果;3)将 OCR 系统由 Google OCR 系统替换成效果更好的 Microsoft OCR 系统,在不增加额外数据的情况下,KR-M4C 模型结果相较于同等条件下的 SA-M4C 模型,在验证集上准确率提升 1.7%(第 13 行);4)增加 ST-VQA 数据集作为训练数据,KR-M4C 模型在验证集和测试集上准确率最好,分别为 49.27% 和 49.63%,比同等情况下的 SA-M4C 分别提升了 1.3% 和 1.1%。

为进一步验证 KR-M4C 方法的有效性,在 ST-VQA 数据集上也进行了对比实验,结果如表 4 所示。可以看出,KR-M4C 方法比现有最好结果在验证集上的准确率提升了 4.7%,在验证集和测试集上的 ANLS 指标均提升了 5%。

此外,为进一步分析 M4C、SA-M4C 和 KR-M4C 模型之间的差异,对算法的模型尺寸和复杂度进行对比,结果如表 5 所示。可以看出:1)6(R)的 KR-M4C 相较 6(N)的 M4C 和 2(N)+4(S)的 SA-M4C,模型参数量和 FLOPs 几乎没有增长,这是因为 KR-M4C 模型在式(6)中引入的模型参数和计算量相比 Transformer 主干网络几乎可以忽略;2)相比 M4C 和 SA-M4C,KR-M4C 在平均推理时间上用时分别增加 48% 和 12%,增加的时间主要用于计算预测答案单词与 OCR 单词之间的余弦相似度。如何优化这部分计算过程以进一步提升方法的计算效率是未来拟开展的重要工作。

3.4 典型样例分析

为了更好地理解 KR-M4C 的表现,本文挑选若干典型样例进行分析,相关结果如图 5 所示。可以

表 4 ST-VQA 数据集上的结果
Table 4 Comparative results on ST-VQA dataset

方法	验证集准确率/%	验证集 ANLS	测试集 ANLS
SAN + STR(Biten 等,2019)	–	–	0.135
M4C(Hu 等,2020)	38.05	0.472	0.462
SMA(Gao 等,2021)	–	–	0.466
CRN(Liu 等,2020)	–	–	0.483
LaAPNet(Han 等,2020)	39.74	0.497	0.485
SA-M4C(Kant 等,2020)	42.23	0.512	0.504
KR-M4C(本文)	46.99	0.562	0.555

注:加粗字体表示各列最优结果,“–”表示无实验数据。

表 5 模型复杂性的对比结果
Table 5 Comparative results of model complexity

方法	网络参数量/M	FLOPs/G	推理单幅图像平均耗时/ms
M4C(Hu 等,2020)	96.63	18.3	89
SA-M4C(Kant 等,2020)	96.63	18.4	118
KR-M4C(本文)	96.65	18.4	132



Q:What should I“made?”?
A:Technology

M4C: the first ×
SA-M4C: the time ×
KE-M4C(本文): technology √



Q:What kind of cognac is this?
A:Super premium

M4C: cognac ×
SA-M4C: super ×
KE-M4C(本文):super premium √



Q:What are the numbers on top?
A:22268888

M4C: 0of6e ×
SA-M4C: 0of6e ×
KE-M4C(本文): 22268888 √



Q:What is the first word on the top left of the boy's t-shirt on the left?
A:Racing

M4C: formula ×
SA-M4C: racing √
KE-M4C(本文): racing √



Q:What is written on the front of that building?
A:Lion gate

M4C: not a question ×
SA-M4C: lion gate √
KE-M4C(本文): lion gate √



Q:What food does the dominic's food truck primarily sell?
A:Italian sausage

M4C: Italian ×
SA-M4C: Italian ×
KE-M4C(本文): Italian sausage √



Q:What is the largest measurement we can see on this ruler?
A:50

M4C: 40 ×
SA-M4C: 40 ×
KE-M4C(本文): 40 ×



Q:What brand is this beverage?
A:Fayrouz

M4C: ayrouz ×
SA-M4C: ayrouz ×
KE-M4C(本文): ayrouz ×

图 5 M4C、SA-M4C 和 KR-M4C 模型的预测结果示例

Fig.5 Examples of the predictions of M4C, SA-M4C and KR-M4C models

看出,1)KR-M4C 相比 M4C 和 SA-M4C 获得了总体上更好的结果,体现了引入知识增强后模型对场景文本理解能力的提升。2)SA-M4C 和 KR-M4C 在涉

及相对空间关系时的表现相比 M4C 具有明显优势,可以准确回答“What is the first word on the top left of the boy’s t-shirt on the left?” 这样需要复杂空间推

理才能准确回答的任务。3) 得益于 KR-M4C 建立的语义关联知识, 模型可以发现预测答案与 OCR 单词间隐含的语义关联, 因此在涉及多个单词的答案时表现比只考虑空间关联的 SA-M4C 方法表现更好。4) 面对部分场景信息确实, 需要“联想”能力才能理解的复杂场景时, 所有方法均表现不佳。这反映了现有场景文本视觉问答框架的性能瓶颈, 有待后续更深入的研究。

4 结 论

本文提出一种融合知识表征的多模态 Transformer 的场景文本视觉问答方法, 在基线方法 M4C 基础上引入“空间关联”和“语义关联”两种互补的先验知识, 提出知识表征增强的 KR-M4C 模型, 实现两种知识与多模态数据的统一建模表达。本文在 TextVQA 和 ST-VQA 两个常用的场景文本视觉问答数据集上进行实验验证, 相比现有最好方法取得了明显的性能提升。本文提出的知识表征增强的多模态 Transformer 框架具有通用性, 除了应用于场景文本视觉任务, 也为其他相关多模态学习任务的方法改进提供了平台。

相比现有方法, 本文提出的 KR-M4C 方法带来了显著的性能提升, 但其性能受限于外部的 OCR 系统的识别能力。如何在模型中引入 OCR 识别模块, 进行端到端的联合优化是未来一个有意义的研究方向。此外, 现有方法的性能与人工标注的训练样本数量紧密相关。如何突破这种标注数据制约, 利用天然的弱标注数据实现模型的预训练以支撑更大更深模型的有效训练, 进一步提升模型的表达能力也是值得探索的重要方向。

参考文献 (References)

- Almazán J, Gordo A, Fornés A and Valveny E. 2014. Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (12): 2552-2566 [DOI: 10.1109/TPAMI.2014.2339814]
- Anderson P, Fernando B, Johnson M and Gould S. 2016. SPICE: semantic propositional image caption evaluation//*Proceedings of the 14th European Conference on Computer Vision*. Amsterdam, the Netherlands: Springer: 382-398 [DOI: 10.1007/978-3-319-46454-1_24]
- Antol S, Agrawal A, Lu J S, Mitchell M, Batra D, Zitnick C L and Parikh D. 2015. VQA: visual question answering//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Santiago, Chile: IEEE: 2425-2433 [DOI: 10.1109/ICCV.2015.279]
- Ben-Younes H, Cadene R, Cord M and Thome N. 2017. MUTAN: multimodal tucker fusion for visual question answering//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE: 2612-2620 [DOI: 10.1109/ICCV.2017.285]
- Biten A F, Tito R, Mafla A, Gomez L, Rusiñol M, Jawahar C V, Valveny E and Karatzas D. 2019. Scene text visual question answering//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul, Koera (South): IEEE: 4291-4301 [DOI: 10.1109/ICCV.2019.00439]
- Bojanowski P, Grave E, Joulin A and Mikolov T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135-146 [DOI: 10.1162/tacl_a_00051]
- Borisyuk F, Gordo A and Sivakumar V. 2018. Rosetta: large scale system for text detection and recognition in images//*Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. London, UK: ACM: 71-79 [DOI: 10.1145/3219819.3219861]
- Chen Y C, Li L J, Yu L C, El Kholy A, Ahmed F, Gan Z, Cheng Y and Liu J J. 2020. UNITER: universal image-text representation learning//*Proceedings of the 16th European Conference on Computer Vision*. Glasgow, UK: Springer: 104-120 [DOI: 10.1007/978-3-030-58577-8_7]
- Cui Y H, Yu Z, Wang C Q, Zhao Z Z, Zhang J, Wang M and Yu J. 2021. ROSITA: enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration//*Proceedings of the 29th ACM International Conference on Multimedia*. Chengdu, China: ACM: 797-806 [DOI: 10.1145/3474085.3475251]
- Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//*Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, USA: IEEE: 248-255.
- Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//*Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]
- Fukui A, Park D H, Yang D, Rohrbach A, Darrell T and Rohrbach M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding//*Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, USA: Association for Computational Linguistics: 457-468 [DOI: 10.18653/v1/D16-1044]
- Gao C Y, Zhu Q, Wang P, Li H, Liu Y L, Van Den Hengel A and Wu

- Q. 2021. Structured multimodal attentions for textvqa [EB/OL]. [2021-11-26]. <https://arxiv.org/pdf/2006.00753.pdf>
- Gao D F, Li K, Wang R P, Shan S G and Chen X L. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 12746-12756 [DOI: 10.1109/CVPR42600.2020.01276]
- Gurari D, Li Q, Stangl A J, Guo A H, Lin C, Grauman K, Luo J B and Bigham J P. 2018. VizWiz grand challenge: answering visual questions from blind people//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 3608-3617 [DOI: 10.1109/CVPR.2018.00380]
- Han W, Huang H T and Han T. 2020. Finding the evidence: localization-aware answer prediction for text visual question answering//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain; International Committee on Computational Linguistics: 3118-3131 [DOI: 10.18653/v1/2020.coling-main.278]
- Hu H, Gu J Y, Zhang Z, Dai J F and Wei Y C. 2018. Relation networks for object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 3588-3597 [DOI: 10.1109/CVPR.2018.00378]
- Hu R H, Singh A, Darrell T and Rohrbach M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textVQA//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 9989-9999 [DOI: 10.1109/CVPR42600.2020.01001]
- Jiang Y, Natarajan V, Chen X L, Rohrbach M, Batra D and Parikh D. 2018. Pythia v0.1: the winning entry to the VQA challenge 2018 [EB/OL]. [2021-11-28]. <https://arxiv.org/pdf/1807.09956.pdf>
- Kant Y, Batra D, Anderson P, Schwing A, Parikh D, Lu J S and Agrawal H. 2020. Spatially aware multimodal transformers for textVQA//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK; Springer: 715-732 [DOI: 10.1007/978-3-030-58545-7_41]
- Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar V R, Lu S J, Shafait F, Uchida S and Valveny E. 2015. ICDAR 2015 competition on robust reading//Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia; IEEE: 1156-1160 [DOI: 10.1109/ICDAR.2015.7333942]
- Karpathy A and Li F F. 2015. Deep visual-semantic alignments for generating image descriptions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA; IEEE: 3128-3137 [DOI: 10.1109/CVPR.2015.7298932]
- Kim J H, Jun J and Zhang B T. 2018. Bilinear attention networks [EB/OL]. [2021-05-21]. <https://arxiv.org/pdf/1805.07932.pdf>
- Kim J H, On K W, Lim W, Kim J, Ha J W and Zhang B T. 2017. Hadamard product for low-rank bilinear pooling//Proceedings of the 5th International Conference on Learning Representations. Toulon, France; OpenReview
- Krishna R, Zhu Y K, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L J, Shamma D A, Bernstein M S and Li F F. 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123 (1): 32-73 [DOI: 10.1007/s11263-016-0981-7]
- Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Mallocci M, Kolesnikov A, Duerig T and Ferrari V. 2020. The open images dataset v4. International Journal of Computer Vision, 128 (7): 1956-1981 [DOI: 10.1007/s11263-020-01316-z]
- Lee K H, Chen X, Hua G, Hu H D and He X D. 2018. Stacked cross attention for image-text matching//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany; Springer: 212-228 [DOI: 10.1007/978-3-030-01225-0_13]
- Liu F, Xu G H, Wu Q, Du Q, Jia W and Tan M K. 2020. Cascade reasoning network for text-based visual question answering//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA; ACM: 4060-4069 [DOI: 10.1145/3394171.3413924]
- Lu J S, Batra D, Parikh D and Lee S. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [EB/OL]. [2021-08-06]. <https://arxiv.org/pdf/1908.02265.pdf>
- Mishra A, Alahari K and Jawahar C V. 2013. Image retrieval using textual cues//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia; IEEE: 3040-3047 [DOI: 10.1109/ICCV.2013.378]
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z M, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J J and Chintala S. 2019. PyTorch: an imperative style, high-performance deep learning library//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada; NIPS: 8026-8037
- Pennington J, Socher R and Manning C D. 2014. GloVe: global vectors for word representation//Proceedings of 2014 conference on Empirical Methods in Natural Language Processing. Doha, Qatar; ACL: 1532-1543 [DOI: 10.3115/v1/D14-1162]
- Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada; NIPS: 91-99
- Singh A, Natarajan V, Shah M, Jiang Y, Chen X L, Batra D, Parikh D and Rohrbach M. 2019. Towards VQA models that can read//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE: 8317-8326 [DOI:

- 10.1109/CVPR.2019.00851]
- Tan H and Bansal M. 2019. LXMERT: learning cross-modality encoder representations from transformers//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics: 5100-5111 [DOI: 10.18653/v1/D19-1514]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: NIPS: 6000-6010
- Veit A, Madera T, Neumann L, Matas J and Belongie S. 2016. COCO-Text: dataset and benchmark for text detection and recognition in natural images [EB/OL]. [2021-01-26]. <https://arxiv.org/pdf/1601.07140.pdf>
- Xie S N, Girshick R, Dollár P, Tu Z W and He K M. 2017. Aggregated residual transformations for deep neural networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5987-5995 [DOI: 10.1109/CVPR.2017.634]
- Yao T, Pan Y W, Li Y H and Mei T. 2018. Exploring visual relationship for image captioning//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 711-727 [DOI: 10.1007/978-3-030-01264-9_42]
- Yu L C, Tan H, Bansal M and Berg T L. 2017a. A joint speaker-listener-reinforcer model for referring expressions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3521-3529 [DOI: 10.1109/CVPR.2017.375]
- Yu Z, Yu J, Cui Y H, Tao D C and Tian Q. 2019. Deep modular co-attention networks for visual question answering//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6274-6283 [DOI: 10.1109/CVPR.2019.00644]

- Yu Z, Yu J, Fan J P and Tao D C. 2017b. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1839-1848 [DOI: 10.1109/ICCV.2017.202]
- Zhou B L, Tian Y D, Sukhbaatar S, Szlam A and Fergus R. 2015. Simple baseline for visual question answering [EB/OL]. [2021-12-07]. <https://arxiv.org/pdf/1512.02167.pdf>

作者简介



余宙, 1988年生, 男, 副教授, 硕士生导师, 主要研究方向为多模态数据分析与推理。
E-mail: yuz@hdu.edu.cn



俞俊, 通信作者, 男, 教授, 博士生导师, 主要研究方向为多媒体分析与检索。
E-mail: yujun@hdu.edu.cn

朱俊杰, 男, 硕士研究生, 主要研究方向为多媒体分析与视觉问答。E-mail: zhujj@hdu.edu.cn
匡振中, 男, 副教授, 主要研究方向为多媒体隐私保护。
E-mail: zzkuang@hdu.edu.cn