软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):1997−1999 [doi: 10.13328/j.cnki.jos.006776]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合预训练技术的多模态学习研究专题前言*

宋雪萌[1], 聂礼强[2], 申恒涛[3], 田 奇[4], 黄 华[5]

[1](山东大学 计算机科学与技术学院, 山东 青岛 266237)

[2](哈尔滨工业大学 (深圳) 计算机科学与技术学院, 广东 深圳 518055)

[3](电子科技大学 计算机科学与工程学院, 四川 成都 611731)

[4](华为技术有限公司, 广东 深圳 518129)

[5](北京师范大学 人工智能学院, 北京 100875)

通信作者: 宋雪萌, E-mail: songxuemeng@sdu.edu.cn

中文引用格式: 宋雪萌, 聂礼强, 申恒涛, 田奇, 黄华. 融合预训练技术的多模态学习研究专题前言. 软件学报, 2023, 34(5): 1997–1999. http://www.jos.org.cn/1000-9825/6776.htm

在当今信息爆炸的时代, 海量多媒体数据涌现在互联网上. 为更好地理解和分析多媒体数据, 多模态学习逐渐成为研究热点. 深度学习凭借其优秀的数据表征能力, 成为多模态学习的主要技术之一. 然而, 现有的有标注数据集规模有限, 往往难以保证复杂深度学习模型的泛化能力, 这给传统的多模态学习研究带来了巨大挑战. 为此, 预训练技术逐渐引发国内外诸多学者的关注, 为多模态学习研究领域提供了新的发展机遇. 专题强调多模态学习与预训练技术的深度融合, 研究融合预训练技术的多模态学习, 包括两方面: (1) 利用预训练模型强大的通用表征能力, 解决多模态学习领域研究相关的痛点、难点问题; (2) 利用多模态学习领域丰富的理论积淀, 促进预训练相关技术的发展. 专题围绕多模态数据分析技术, 通过探讨多模态学习与预训练技术的深度融合, 重点关注面向多模态学习的预训练技术、融入预训练技术的多模态内容理解以及融入预训练技术的多模态生成, 旨在有效结合多模态学习技术与预训练技术, 实现多模态学习技术和预训练技术的相辅相成, 促进多模态学习领域的研究发展.

本专题公开征文, 共收到投稿 41 篇. 论文均通过了形式审查, 内容涉及融合预训练技术的多模态学习研究. 特约编辑先后邀请了 40 多位专家参与审稿工作, 每篇投稿至少邀请 2 位专家进行评审. 稿件经初审、复审、ChinaMM2022 会议宣读和终审 4 个阶段, 历时 5 个月, 最终有 10 篇论文入选本专题. 根据主题, 这些论文可以分为 3 组.

(1) 面向多模态学习的预训练技术

《视觉语言预训练综述》系统梳理了 2019 年以来视觉语言预训练相关工作的发展脉络, 介绍了常用的视觉语言预训练技术和相关预训练数据集, 同时比较了不同视觉语言预训练模型在不同任务下、不同数据集上的性能表现.

《面向视觉语言理解与生成的多模态预训练方法》提出了视觉语言理解和生成统一的多模态预训练技术. 该技术拓展了现有的预训练范式, 同时使用了随机掩码和因果掩码, 使得预训练模型能够具有自回归的生成能力.

(2) 融入预训练技术的多模态内容理解

《基于虚拟属性学习的文本-图像行人检索方法》融入预训练技术, 提出了一种基于虚拟属性学习的文本-图像行人检索方法. 该方法基于行人属性不变性和跨模态语义一致性进行属性解耦, 并通过行人身份标签引导属性解耦过程. 同时设计了一个基于语义推理的特征学习模块, 通过图模型增强特征的跨模态识别能力.

《预训练驱动的多模态边界感知视觉 Transformer》面向图像篡改检测任务, 提出了一种预训练驱动的多模态边界感知方法. 该方法通过结合 Transformer 模块与预训练技术, 提取全局上下文信息, 并使用边界感知模块和渐进式语义生成模块探索空间和通道方面的相关性, 逐级生成检测结果图.

* 收稿时间: 2022-09-23; jos 在线出版时间: 2022-09-23

《多模态引导的局部特征选择小样本学习方法》将预训练技术和多模态学习技术引入小样本学习领域, 提出了多模态引导的局部特征选择小样本学习方法. 该方法包含模型预训练和小样本元学习两个阶段, 可达到更好的小样本学习效果.

《基于自监督图对比学习的视频问答方法》面向视频问答任务提出了一种基于自监督图对比学习的通用框架, 通过生成不同的原始输入图的随机残缺子图来提升模型的泛化能力. 同时设计了一种基于图卷积网络的掩码机制, 包含节点丢弃 (Node-dropping) 和边丢弃 (Edge-dropping) 两种操作, 提升了模型的抗扰动能力.

(3) 融入预训练技术的多模态生成

《基于多级残差映射器的文本驱动人脸图像生成和编辑》提出了一个统一的文本驱动人脸图像生成和编辑框架. 该框架基于对比语言-图像预训练模型 CLIP 和预训练 StyleGAN2 模型, 为不同语义级别特征学习残差映射器, 将输入条件映射为潜在编码, 从而实现文本驱动人脸图像生成和编辑功能.

《基于多域 VQGAN 的文本生成国画方法研究》提出了一种多域文生成图模型, 能够用一个模型生成多个域的图像. 通过融入预训练技术, 该方法无需大量标注数据就可实现高质量的多域图像生成.

《预训练模型特征提取的双对抗磁共振图像融合网络研究》提出了基于预训练的双对抗图像融合模型. 该方法采用预训练的特征提取模型, 引入其中的先验知识, 提取医学图像特征, 并设计了基于对抗学习的特征融合模型, 实现医学图像融合.

《基于视觉区域聚合与双向协作的端到端图像描述生成》引入 Transformer 框架提取图像网格型视觉特征, 指导模型生成的文本涵盖图像全局语义; 同时设计视觉区域特征聚合模块, 提取图像区域型视觉特征, 指导模型生成的文本能够描述图像细粒度语义.

本专题主要面向多媒体、计算机视觉、自然语言处理等领域的研究人员和工程人员, 反映了我国学者在融合预训练技术的多模态学习领域最新的研究. 感谢《软件学报》编委会对专题工作的指导和帮助, 感谢专题全体评审专家及时、耐心、细致的评审工作, 感谢踊跃投稿的所有作者. 希望本专题能够对融合预训练技术的多模态学习相关领域的研究工作有所促进.

宋雪萌(1990－), 女, 博士, 山东大学副教授, 博士生导师, 在中国计算机学会推荐的国内外著名学术期刊与会议上发表论文 50 余篇, 出版学术专著 3 部. 入选“人工智能相关领域全球女性学者”名单和全球华人 AI 青年学者榜单. 获山东省科学技术进步奖一等奖. 主要研究领域为信息检索和多媒体计算.

聂礼强(1985－), 男, 博士, 哈尔滨工业大学 (深圳) 教授, 院长, 博士生导师, IAPR Fellow, AAIA Fellow, 发表 CCF A 类论文百余篇, 出版中英文专著 5 部, 谷歌引用 1.6 万余次. 获 SIGIR/SIGMM 最佳论文提名奖、SIGMM Rising Star、达摩院青橙奖、SIGIR 最佳学生论文奖、省科技进步一等 (排 1) 奖. 主要研究领域为多媒体内容分析与搜索.

申恒涛(1977－), 男, 博士, 电子科技大学计算机科学与工程学院院长, 欧洲科学院外籍院士, ACM Fellow, IEEE Fellow, OSA Fellow, 发表了 360 余篇高水平同行评审论文, 包括 150 多篇 IEEE/ACM Transactions 和 250 多篇 CCF A 类论文, 获得了 8 个国际会议和期刊的最佳论文奖. 主要研究领域为多媒体搜索, 计算机视觉, 人工智能.

田奇(1970－), 男, 博士, 华为云人工智能领域首席科学家, 博士生导师, 国际欧亚科学院院士, IEEE Fellow, 海外杰青, 中国科学院海外评审专家. 发表文章 650 余篇, 谷歌引用 4.2 万余次. 曾获 Google Faculty Research Award、UTSA 校长杰出研究奖、中国人工智能学会吴文俊人工智能杰出贡献奖以及多媒体领域 10 大最具影响力的学者, 担任国家自然科学基金会评专家. 主要研究领域为计算机视觉.

黄华(1975－), 男, 博士, 北京师范大学教授, 博士生导师, 国家杰出青年基金获得者, 入选国家高层次人才特殊支持计划 (万人计划) 领军人才, 曾获中国青年科技奖、ICML Outstanding Paper Awards 和 EURASIP Best Paper Award. 主要研究领域为图像/视频处理.