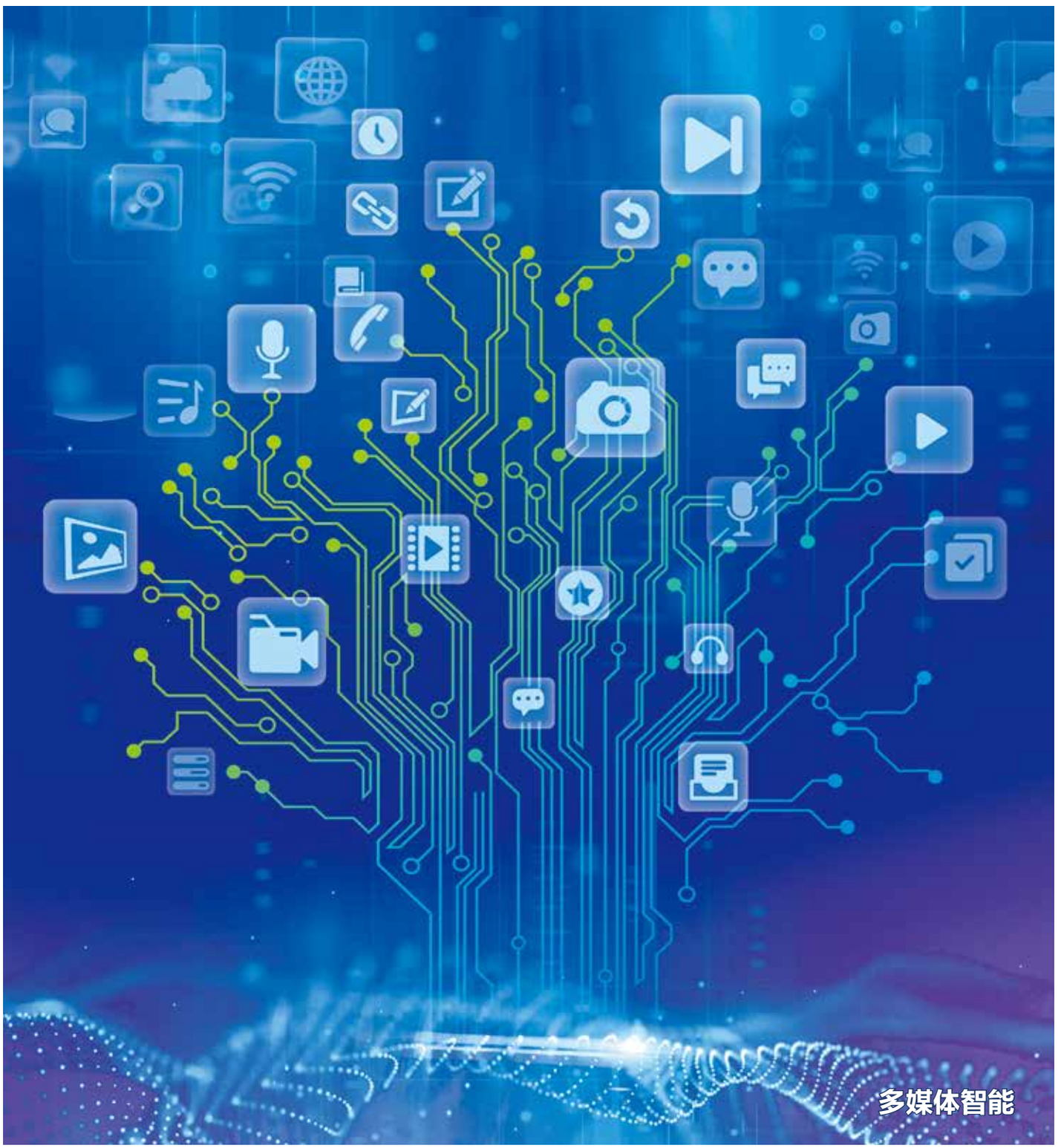
JOURNAL OF IMAGE AND GRAPHICS

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



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

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

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

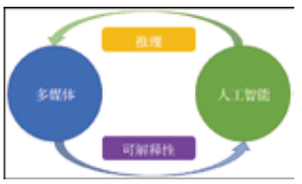
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

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

2022年9月 第27卷 第9期（总第317期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次



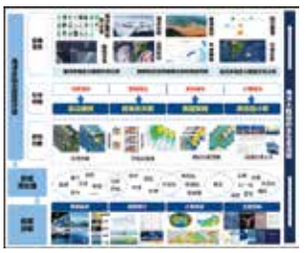
多媒体智能：当多媒体遇到人工智能(第2551页)

《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



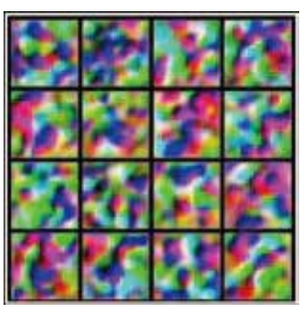
面向海洋的多模态智能计算：挑战、进展和展望(第2589页)

基于深度学习的人—物交互关系检测综述
廖越，李智敏，刘偲 …… 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



基于真实数据感知的模型功能窃取攻击(第2721页)

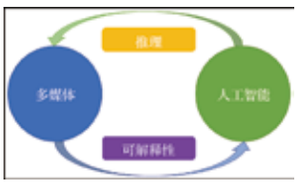
利用时空特征编码的单目标跟踪网络
王蒙蒙，杨小倩，刘勇 …… 2733
结合时空一致性的FairMOT跟踪算法优化
彭嘉淇，王涛，陈柯安，林巍峣 …… 2749

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

Volume 27, Number 9
Published September 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



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

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

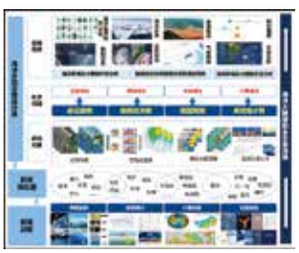
Review

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



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

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

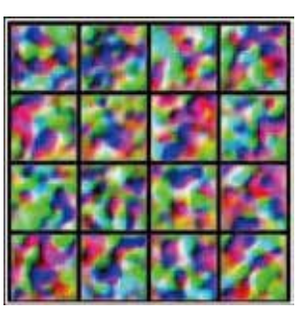
Spatio-temporal consistency based FairMOT tracking algorithm optimization
Peng Jiaqi, Wang Tao, Chen Kean, Lin Weiyao ····· 2749

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



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

中图法分类号:TP181　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)09-2551-23

论文引用格式: Zhu W W, Wang X, Tian Y H and Gao W. 2022. Multimedia intelligence: the convergence of multimedia and artificial intelligence. Journal of Image and Graphics,27(09):2551-2573(朱文武,王鑫,田永鸿,高文. 2022. 多媒体智能:当多媒体遇到人工智能. 中国图象图形学报,27(09):2551-2573)[DOI:10.11834/jig.220086]

多媒体智能:当多媒体遇到人工智能

朱文武[1],王鑫[1],田永鸿[2*],高文[2]

1. 清华大学计算机系,北京　100084; 2. 北京大学计算机学院, 北京　100871

摘　要:　过去10年中涌现出大量新兴的多媒体应用和服务,带来了很多可以用于多媒体前沿研究的多媒体数据。多媒体研究在图像/视频内容分析、多媒体搜索和推荐、流媒体服务和多媒体内容分发等方向均取得了重要进展。与此同时,由于在深度学习领域所取得的重大突破,人工智能(artificial intelligence,AI)在20世纪50年代被正式视为一门学科之后,迎来了一次"新"的发展浪潮。因此,一个问题就自然而然地出现了:当多媒体遇到人工智能时会带来什么?为了回答这个问题,本文通过研究多媒体和人工智能之间的相互影响引入了多媒体智能的概念。从两个方面探讨多媒体与人工智能之间的相互影响:一是多媒体促使人工智能向着更具可解释性的方向发展;二是人工智能反过来为多媒体研究注入了新的思维方式。这两个方面形成了一个良性循环,多媒体和人工智能在其中不断促进彼此发展。本文对相关研究及进展进行了讨论,并围绕值得进一步探索的研究方向分享见解。希望可以对多媒体智能的未来发展带来新的研究思路。

关键词:多媒体技术;人工智能(AI);多媒体智能;多媒体推理;可解释人工智能

Multimedia intelligence: the convergence of multimedia and artificial intelligence

Zhu Wenwu[1], Wang Xin[1], Tian Yonghong[2*], Gao Wen[2]

1. Department of Computer Science and Technology, Tsinghua University, Beijing　100084, China;

2. School of Computer Science, Peking University, Beijing　100871, China

Abstract: Multimedia can be regarded as an integration of various medium such as videos, static images, audios, and texts. Thanks to the rapid development of emerging multimedia applications and services, a huge amount of multimedia data has been generated to advance multimedia research. Furthermore, multimedia research has made great progress in image/video processing and analysis, including search, recommendation, streaming, and content delivery. Since artificial intelligence (AI) became an official academic discipline in the 1 950 s, it has experienced a "new" wave of boost based on deep learning techniques. Its development has been witnessed in the past decades, including expert systems, intelligent search and optimization, symbolic and logical reasoning, probabilistic methods, statistical learning methods, artificial neural networks, etc. As such, a natural question arises: "What will happen when multimedia meets AI?" To answer this question, we introduce the concept of multimedia intelligence by investigating the mutual influences between multimedia and AI. Multimedia drives AI towards a more explainable paradigm, because semantic information is able to enhance the explainability

收稿日期:2022-01-27;**修回日期**:2022-06-29;**预印本日期**:2022-07-06

*通信作者:田永鸿　yhtian@pku.edu.cn

基金项目:科技创新2030-"新一代人工智能"重大项目(2020AAA0106300);国家自然科学基金项目(62222209, 62102222, 62250008)

Supported by: National Key R&D Program of China (2020AAA0106300); National Natural Science Foundation of China (62222209, 62102222, 62250008)

of AI models. At the same time, AI is beneficial for multimedia technology to pocess the advanced ability of reasoning. AI promotes the human-like perception and reasoning processes, which can lead to more inferable multimedia processing and analizing techniques. These mutual influences form a loop in which multimedia and AI interactively enhance each other. To sum up, we discuss the recent advances in literature and share our insights on future research directions deserving further study. We hope this paper can bring new inspirations for future development of multimedia intelligence.

Key words: multimedia technology; artificial intelligence (AI); multimedia intelligence; multimedia reasoning; explainable artificial intelligence

0 引 言

从20世纪60年代第1次出现到今天的广泛使用,多媒体一词一直有着不同的含义。现在称多媒体为“交互式访问的多种媒体组合,包括视频、静态图像、音频和文本。”(“an electronically delivered combination of media including videos, still images, audios, and texts in such a way that can be accessed interactively.”)(https://en.wikipedia.org/wiki/Multimedia)经过二十多年的迅猛发展(Li 等,2013;Zhang 和 Rui,2013),多媒体研究在图像/视频内容分析、多媒体搜索和推荐、流媒体和多媒体内容分发等方面取得了很大进展(Zhu 等,2015)。人工智能理论早在20世纪50年代之前就出现在了学术研究者的视野中,并在几十年内发展出了各种方法,包括专家系统、智能搜索与优化、符号与逻辑推理、概率论方法、统计学习方法和人工神经网络等。多媒体和人工智能这两个重要的研究领域在以前几乎是各自独立发展,直到各种丰富的多媒体数据不断增加,人工智能才得以发展出更多实用模型来处理各种真实世界的多媒体信息,进而在真实世界的场景中得到应用。因此,一个值得深入思考的关键问题出现了:当多媒体和人工智能结合在一起时会带来什么?

为了回答这个问题,本文通过探索多媒体和人工智能之间的相互影响,提出了多媒体智能的概念。当多媒体作用于人工智能时,多媒体促使人工智能向着更具可解释性的方向发展:在丰富的、带有可解释语义信息数据的帮助下,大量多媒体数据为人工智能模型可解释能力的增强带来了可能。由此产生的新一轮人工智能热潮也可以从国内外顶尖大学或中央政府为未来人工智能制定的许多计划中看出来。例如,美国斯坦福大学在2014年为人工智能提出了“人工智能100年(AI 100)”计划,以理解人们是如何工作、生活和娱乐的。此外,美国政府在2016年宣布了一项“为人工智能的未来做准备”的提案,成立了“人工智能和机器学习委员会”。欧盟(European Union, EU)提出了欧洲人工智能新理论,强调让以人为中心的人工智能可信任,包括技术鲁棒性、安全性、透明性和可靠性等。与此同时,中国也制定了新一代人工智能发展计划,强调可解释的、可推理的人工智能。当多媒体被人工智能所影响时,人工智能反过来会带来更加可推理的多媒体技术,这相较于现有的多媒体技术方法具有更强的解释性并且更符合人类认知习惯的推理能力。人工智能的终极目标之一是弄清楚一个智能体如何才能在现实世界中成长与发展,并再现这一过程。感知和推理能力是使得人类能在多样环境中生存的一个重要因素。因此,研究人工智能中的类人感知和推理过程将带来具有感知和推理能力的可推断的多媒体,能够对感知到的外部环境进行推理并做出决策。然而,关注这一方向(即利用人工智能来增强多媒体的推理能力、推动多媒体发展)的工作却很少。本文从两个方面探讨多媒体与人工智能之间的相互关系和相互影响:

1)多媒体对人工智能的影响:多媒体推动人工智能向着更具可解释性的方向发展。

2)人工智能对多媒体的影响:人工智能促进多媒体技术推断能力的发展。

因此,本文将多媒体智能定义为多媒体和人工智能的融合。如图1所示,它的内涵是多媒体和人工智能相互促进和增强的良性循环,它的特点在于具有可解释性和推断能力。

更具体地说,由于当前蓬勃发展的机器学习理论方法在基于人工智能的数据建模和分析中占据了统治地位,同时机器学习技术作为目前人工智能领域中的代表广泛应用于多媒体领域,因此为便于叙述和阐释,以机器学习为例,从本文下面两个方向讨论多媒体与人工智能之间的双向影响:

1)多媒体信息通过催生许多适用于特定多媒

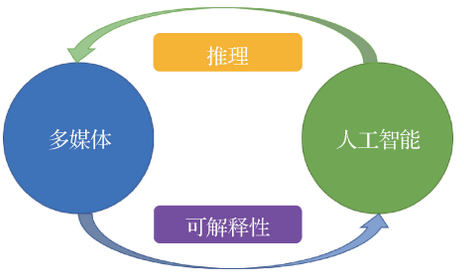


图 1　多媒体智能的“循环”

Fig. 1　Cycle of multimedia intelligence

体任务且更易于解释的机器学习技术来促进机器学习发展，并扩大了机器学习的应用范围。

2）机器学习赋予多媒体技术更强大的推理能力，增强了多媒体信息分析模型的推断能力。

本文对多媒体智能的相关研究工作及进展进行讨论与总结，同时指出还有哪些值得研究的工作尚未完成，以及如何才能完成。此外，还阐述了可能对多媒体智能产生深远影响且具有前景的研究方向，并提出了一些不成熟的见解。

1　多媒体推动机器学习发展

一方面，多媒体数据的多模态核心推动机器学习发展出了许多新兴技术，来帮助多媒体很好地捕捉和建模多媒体数据的异构特征（Cord 和 Cunningham，2008）。另一方面，大量的多媒体数据使得视听语音识别、图像/视频标注和视觉问答等多种多模态应用成为了可能。如图 2 所示，本节将从两个方面讨论多媒体促进机器学习发展的方式：多媒体如何推动机器学习技术发展；多媒体如何推动机器学习应用发展。

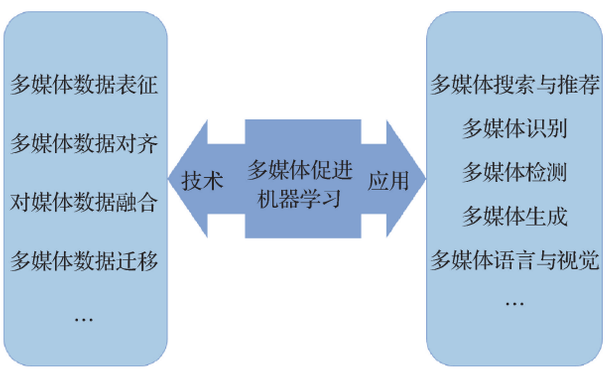


图 2　多媒体推动机器学习发展

Fig. 2　Multimedia promots machine learning

1.1　多媒体推动机器学习技术发展

多媒体数据包含图像、音频和视频等多种类型的数据，其中单模态数据在过去十年中被研究者广泛研究。然而，近来多模态和异构的多媒体数据越来越多，这给机器学习算法在正确处理多模态数据时试图精确捕捉不同模态之间的关系带来了巨大的挑战。因此，本文聚焦在多模态多媒体数据上，总结了多模态多媒体数据分析中的 4 个基本问题，即多媒体数据表征、多媒体数据对齐、多媒体数据融合和多媒体数据迁移，并重点介绍了相应的机器学习技术，旨在通过解决这 4 个问题正确地处理各种多媒体数据。

1）多媒体数据表征。学习多媒体数据表征的方法主要可以分为两类：联合表征和调和表征。联合表征将多个单模态数据组合到同一个特征空间中，而调和表征在将不同模态的数据进行分别处理的同时对它们施加某些相似的约束，使它们在坐标空间中具有可比性。为了获得多媒体数据的联合表征，研究人员设计和利用了元素操作、特征串联、全连接层、多模态深度置信网络（Srivastava 和 Salakhutdinov，2012）、多模态压缩双线性池（Fukui 等，2016）和多模态卷积神经网络（Ma 等，2015）等方法来组合不同模态的数据。为了获得调和表征，一个典型的例子是深度视觉语义嵌入模型（deep visual-semantic embedding model，DeViSE）（Frome 等，2013），它构造了一个从图像到文本特征的简单线性映射，这样相对应的注释和图像表征之间的内积值将大于非对应的值。其他一些工作也在两个单模态自动编码器的共享隐藏层上建立了调和的空间（Wang 等，2015；Yuan 等，2019a）。图 3 展示了多模态数据表征的一个示例。

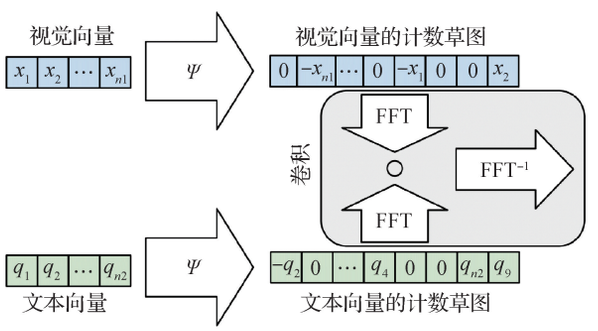


图 3　多模态压缩双线性池（Fukui 等，2016）

Fig. 3　Multimodal compact bilinear pooling（MCB）（Fukui et al.，2016）

越来越多的工作表明了深度自注意力网络(Transformer)在众多领域取得成效,因此使用Transformer来统一多模态的表征成为一种可能。如LXMERT(learning cross-modality encoder representations from Transformers(Tan和Bansal,2019))、Oscar(object-semantics aligned pre-training for vision-language tasks(Li等,2020))等工作,通过Transformer来获取数据的联合表征;而CLIP(contrastive language image pre-training(Radford等,2021))、BriVL(bridging vision and language by large-scale multi-modal pre-training(Huo等,2021))则用以获取数据的调和表征。

2)多媒体数据对齐。多模态多媒体数据对齐是理解多模态数据的一个基本问题,其目的是发现两个或多个模态实例之间的关系和对齐方式。多模态问题,如时域语句定位(Gao等,2017;Hendricks等,2017;Yuan等,2019b)和描述目标定位(Liu等,2019a;Zhang等,2018)都属于多模态数据对齐的研究领域,因为它们需要将句子或短语与相应的视频片段或图像区域对齐。多模式数据对齐可分为两种主要类型——隐式对齐和显式对齐。Baltrušaitis等人(2019)将显式多模态对齐定义为两种或以上模态的实例对齐。而隐式对齐一般用于任务的中间步骤。隐式对齐的模型不直接对齐数据也不依赖于监督数据对齐样本,而是通过模型训练学习如何以隐藏的方式来对齐数据。对于显式对齐,Malmaud等人(2015)利用隐马尔可夫模型(hidden Markov model,HMM)将菜谱步骤与(自动生成的)烹饪视频字幕对齐,Bojanowski等人(2015)通过学习视觉和文本模态之间的线性映射来解决时域对齐问题,以便在视频中自动为句子找到对应的时间(帧)戳。对于隐式对齐,注意力机制(Bahdanau等,2015)作为机器学习领域中一个典型的工具,能够使解码器更多地关注于需要处理的目标子元素,例如图像(Vinyals等,2015)的某个子区域、视频(Yao等,2015;Yu等,2016)中的帧或片段,句子(Bahdanau等,2015)中的单词和音频(Chan等,2016)的片段等。图4展示了多模态数据对齐的一个示例。

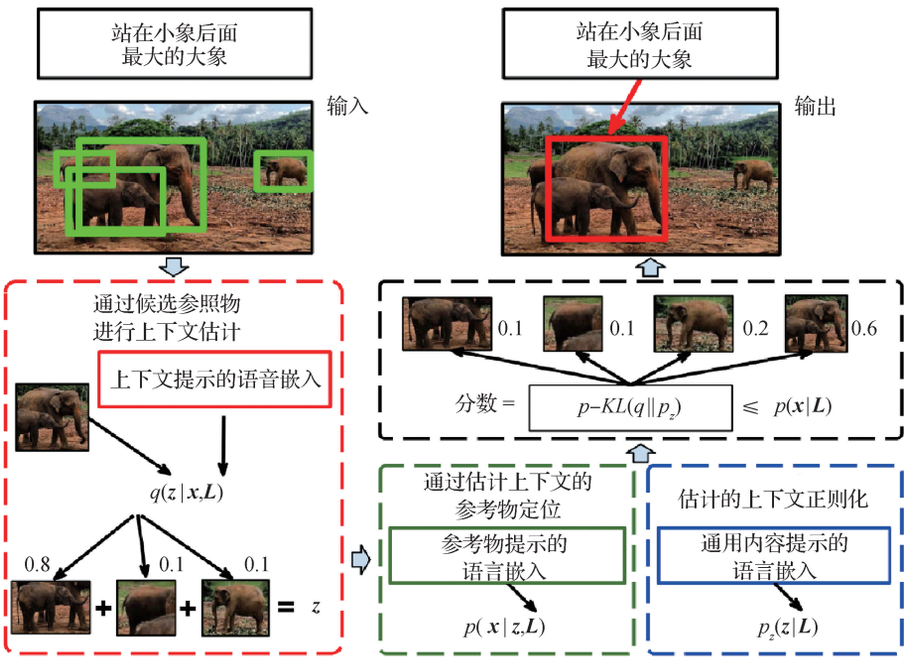
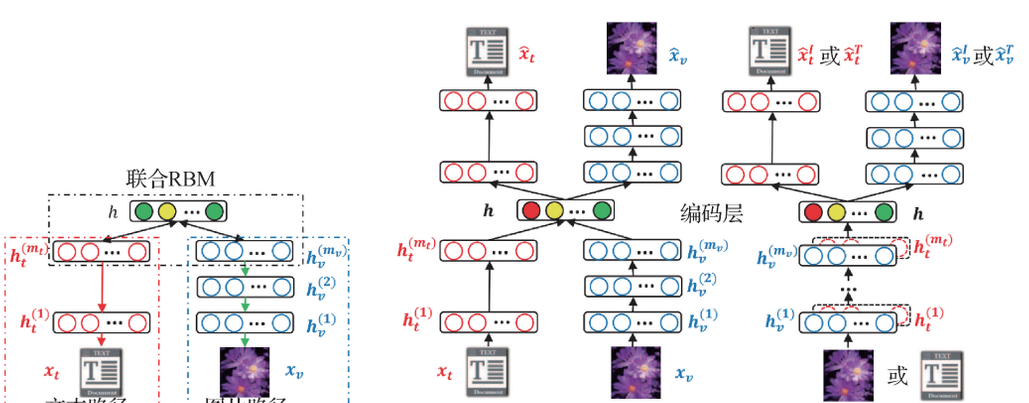


图4 多模态对齐的变分上下文模型(Zhang等,2018)

Fig. 4 The proposed variational context model (Zhang et al., 2018)

3)多媒体数据融合。多模态融合也是多媒体人工智能的关键问题之一。它旨在整合来自多种模式的信号,以预测特定结果:通过分类预测类别(如正或负),或通过回归预测一个连续值(例如,中国某一年的人口)。总体而言,多模态融合方法可以分为两个方向(Baltrušaitis等,2019):与模型无关的方法和基于模型的方法。与模型无关的方法也可以分为3种类型:早融合、晚融合和混合融合。早融合

在提取特征后立即整合来自多种模态的特征(通常通过简单地连接它们的特征来实现)。晚融合在每种模态做出自己对于某个任务的决策(例如,分类或回归)后进行整合。混合融合通过将早融合预测结果和单模态的预测结果通过概率加权聚合在一起来获得合并的输出。与模型无关的方法几乎可以使用任何单模态分类器或回归器来实现,这意味着它们使用的技术不是专门为多模态数据设计的。相比之下,在基于模型的方法中,囊括了 3 类模型来完成多模态融合:基于核的方法、图模型和神经网络。多核学习(multiple kernel learning, MKL)(Gönen 和 Alpaydın,2011)是基于核的支持向量机(support vector machine,SVM)的扩展,允许将不同的核用于来自不同模态/视图的数据。由于核可以看做是估计数据点之间相似性的函数,因此 MKL 中特定于模态的核可以更好地融合异构数据。图模型是多模态融合的又一系列流行方法,可分为生成方法(如耦合(Nefian 等,2002)、阶乘隐马尔可夫模型(Ghahramani 和 Jordan, 1997)以及动态贝叶斯网络(Garg 等,2003))和判别方法(如条件随机场(conditional random field,CRF)(Lafferty 等,2001))。图模型的一个优点是它们能够利用数据的时间和空间结构,使其特别适合用于解决视听语音识别等时域建模的任务。目前,神经网络(Ngiam 等,2011)已广泛用于多模态融合的任务。例如,长短期记忆(long short-term memory, LSTM)网络(Hochreiter 和 Schmidhuber, 1997)已经证明了其在多模态情感序列识别(Wöllmer 等,2013)方面优于图模型的优势;自动编码器(auto encoder, AE)在多模态散列(Wang 等,2015)、多模态量化(Wang 等,2019b)和视频摘要(Yuan 等,2019a)方面达到了令人满意的效果;卷积神经网络(convolutional neural networks, CNN)已广泛用于图像文本检索任务(Ma 等, 2015);CEN(channel exchanging network(Wang 等,2020))则利用通道交换来实现多模态数据的融合,在语义分割和图像转换任务上取得了较好效果;MBT(multimodal bottleneck Transformer(Nagrani 等,2021))通过引入 Transformer,限制不同模态之间信息传递的瓶颈,从而浓缩每个模态中最相关和最必要的信息,在多个视听任务中取得了成功。虽然深度神经网络架构具有从大量数据中学习复杂模式的能力,但它们缺乏推理能力。图 5 展示了多模态融合的一个例子。



(a) 多模态DBN的预训练过程　　(b) 多模态自动编码器的微调过程

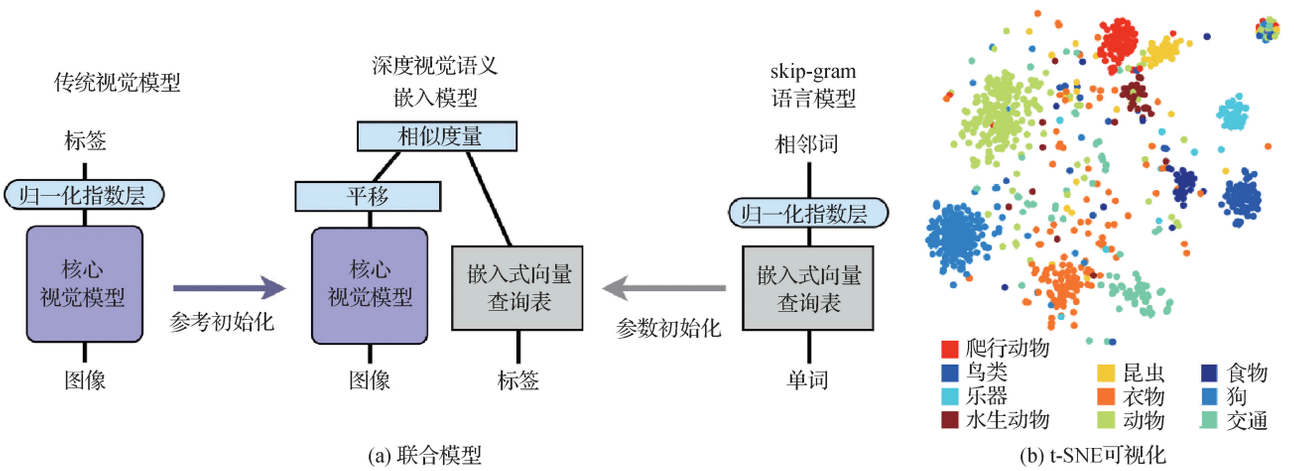
图 5　多模态融合的预训练和微调过程(Wang 等,2015)

Fig. 5　Multimodal fusing in pretraining and fine-tuning(Wang et al. , 2015)

((a) multimodal DBN in pretraining; (b) multimodal AutoEncoder in fine-tuning)

4)多媒体数据迁移。多模态多媒体数据迁移问题旨在不同模态间迁移有用信息,目的是用资源丰富的模态的知识来建模资源贫乏的模态(Baltrušaitis 等,2019)。对于假设模态来自同一数据集,并且实例之间存在直接对应关系的并行多模态数据,迁移学习是实现多模态数据迁移的典型方法。例如,多模态自动编码器(Ngiam 等, 2011; Wang 等,2015)可以通过共享的隐藏层将信息从一个模态迁移到另一个模态,这样不仅能得到合适的多模态表征,而且还能得到更好的单峰特征。迁移学习对于假设模态来自不同的数据集,并且具有重叠的类别或概念而不是重叠的实例的非并行多模态

也是可行的。这种类型的迁移学习通常通过利用调和的多模态表征来实现。例如，DeViSE（Frome 等，2013）将卷积神经网络视觉特征与在单独数据集上训练的词向量（word2vec）文本特征（Mikolov 等，2013）调和起来，以达到利用文本标签来改进分类任务中的图像表征的目的。为了在多模态迁移中处理非并行多模态数据，观念框架（Baroni，2015）和零样本学习（Socher 等，2013）是实际中采用的两种代表性方法。对于实例或概念由第 3 种模态或数据集联系在一起的混合多模态情境（并行和非并行数据的混合），最值得注意的例子是桥关联神经网络（Rajendran 等，2016），它使用一种枢轴模态来学习非并行数据的调和多模态表征。这种方法也可用于机器翻译（Nakov 和 Ng，2009）和音译（Khapra 等，2010），以连接没有并行语料库但共享共同枢轴语言的不同语言。图 6 展示了一个多模态迁移的例子。



(a) 联合模型　　(b) t-SNE可视化

图 6　模型结构与可视化结果（Frome 等，2013）

Fig. 6　Model structure and visualization results (Frome et al., 2013) ((a) joint model; (b) t-SNE visualization)

1.2　多媒体推动机器学习应用发展

如前所述，当前人工智能技术的核心在于机器学习的发展，因此本文将重点介绍几个具有代表性的机器学习应用，包括多媒体搜索与推荐、多媒体识别、多媒体探测、多媒体生成和多媒体语言与视觉，它们因为丰富可用的多模态多媒体数据而得以广泛流行。

1）多媒体搜索与推荐。相似性搜索（Agrawal 等，1993；Ciaccia 等，1997）一直是多媒体信息检索中一个非常基础的研究课题，一个好的相似性搜索策略不仅需要准确性，还需要高效率（Gionis 等，1999）。相似性搜索的经典方法旨在单个模态内搜索相似的内容，例如，在给定文本（图像）作为查询目标的情况下搜索相似的文本（图像）。另外近年来多媒体应用的快速发展创造了大量属于各种不同信息模态的内容，如视频、图像、语音和文本。这些大量的多模态数据带来了对跨多模态内容进行高效、准确的相似性搜索的强烈需求（Rasiwasia 等，2007，2010），例如给定文本搜索相似图像或给定图像搜索相关文本。同时，多模态散列在多媒体检索上的应用和图卷积的结合，可以保证在查询阶段，即使部分模态特征丢失，也能够稳健地捕获各种模态的信息，从而保证搜索的效果（Lu 等，2021）。目前已经有一些关于多模态搜索的研究，更多的详细内容可以从相关的综述论文（Wang 等，2016a；Wang 等，2018）中获取。互联网快速发展带动了各种包含多媒体数据的网络服务的发展，带动了从被动多媒体搜索到主动多媒体搜索的转变，进而带来了多媒体推荐的需求。多媒体推荐技术可以广泛地涵盖用于视频推荐（Yu 等，2019a）、音乐推荐（van den Oord 等，2013）、群组推荐（Wang 等，2016b）和社交推荐（Wang 等，2016c，2017，2019a）等的技术。同时，用户与多模态项目之间的交互也可以作为一种隐式建模，来更好地挖掘多模态内容潜在的信息，从而提高推荐的效果（Zhang 等，2021）。同样，读者可以从相关的综述论文（Zhu 等，2020b）获得关于多模态推荐的更多详细内容。

2）多媒体识别。多媒体研究的最早例子之一是视听语音识别（audio-visual speech recognition，AVSR）（Yuhas 等，1989）。这项工作的灵感来源于

McGurk 效应(McGurk 和 MacDonald,1976),其认为语音感知是在人们的视觉和听觉交互下进行的。McGurk 效应源于一种观察,即人们在观看一位年轻的正在说话的女人的电影时声称听到了音节[da],然而事实是嘴型为[ga]的重复的音节被配音成了[ba]。这项结果激励了许多进行语音研究的研究人员在额外的视觉信息的帮助下扩展他们的方法,特别是来自深度学习方向的研究人员(Afouras 等,2018;Ngiam 等,2011;Petridis 等,2018)。将多模态信息纳入语音识别程序确实提高了识别性能,并在一定程度上提高了可解释性。其他一些人也观察到,当音频信号嘈杂时,视觉信息的优势变得更加突出(Gurban 等,2008;Ngiam 等,2011)。视听语音识别的发展促进了包括视频中的语音增强和识别、视频会议和听觉增强等的广泛应用,特别是在多人嘈杂环境中说话的情况下(Ephrat 等,2018)。尽管之前提及的工作在许多情况下都能够达成较好的效果,但是它们仍然依赖于带标签的训练数据,而在真实生活中的语音数据则多数没有标签,因此半监督(Xu 等,2020b)、自监督(Baevski 等,2020)或无监督(Baevski 等,2021)的语音识别系统能够在语音识别中达到更好的效果。图 7 给出了视听语音识别的流程。

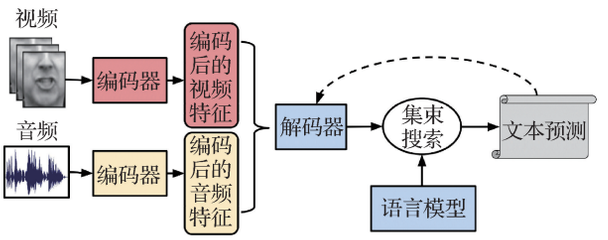


图 7　视听语音识别(AVSR)流程概述(Afouras 等,2018)

Fig. 7　Outline of the audio-visual speech recognition(AVSR) pipeline(Afouras et al. , 2018)

3)多媒体检测。人类活动检测(Ramachandram 和 Taylor,2017)是需要大量利用多媒体数据的一个重要研究领域。由于人类在社交活动中经常表现出高度复杂的行为,因此机器学习算法在理解和识别人类活动时自然需要多模态数据的帮助。深度多模态融合的一些工作通常涉及视觉、听觉、深度、运动甚至骨骼等信息(Chen 等,2015;Escalera 等,2015;Ofli 等,2013)。基于多模态深度学习的方法已应用于有关人类活动的各种任务(Ramachandram 和 Taylor,2017),包括动作检测(Natarajan 等,2012;Singh 等,2016;Xu 等,2020a;Xu,2021)(一个行为可能由多个较短的一系列动作组成)、视线方向估计(Lian 等,2019;Mukherjee 和 Robertson,2015;Zhang 等,2020)、手势识别(Chen 等,2021;Neverova 等,2016;Wu 等,2016)、情感识别(Kahou 等,2016;Poria 等,2016)和面部识别(Ding 和 Tao,2015;Guo 等,2020a;Meng 等,2021;Zhang 等,2015)。现在包含至少 10 个传感器的移动智能手机已经流行起来,催生了许多涉及多模态数据的新应用,包括连续生物识别身份验证(Schultz 和 Sartini,2016;Sitová 等,2016)。图 8 展示了一个多模态检测的例子。

4)多媒体生成。多模态多媒体数据生成是多媒体人工智能的另一个重要方向。给定一个模态中的实体,多媒体生成的任务是生成另一个不同模态中的同一实体。例如,图像/视频标注和从自然语言生成图像/视频是两组典型的应用。多模态生成的核心思想是将信息从一种模态转换为另一种模态,以便在新模态中生成内容。多模态生成中有很多方法,并且通常是基于特定模态的,它们可以分为两种主要类型:基于实例和基于生成(Baltrušaitis 等,2019)。基于实例的方法在模态之间通过构造字典来进行翻译,而基于生成的方法构造能够完成模态间转换的模型。图像转文字(Im2text)(Ordonez 等,2011)是一种典型的基于实例的方法,它利用全局图像表征来检索说明文字并将它们从数据集转移到待查询的图像。其他一些基于实例的方法采用整数线性规划(integer linear programming,ILP)作为优化框架(Kuznetsova 等,2012),该框架搜索用于描述视觉上相似的图像的现有人工组成的短语,然后有选择地组合这些短语以生成待查询图像的新描述。对于基于生成的方法,基于端到端训练的编码器—解码器神经网络设计是目前最流行的多模态生成技术之一。这些模型背后的主要思想是首先将源模态编码为压缩的矢量表示,然后使用解码器生成目标模态。虽然编码器—解码器模型首先被用于机器翻译(Kalchbrenner 和 Blunsom,2013;Sutskever 等,2014),但它们后来被进一步用于解决图像/视频标注(Kuznetsova 等,2021;Venugopalan 等,2015;Vinyals 等,2015)和图像/视频/语音生成(Liu 等,2019d;Mansimov 等,2016;Owens 等,2016;Ramesh 等,2021,2022;Reed 等,2016;van den Oord 等,2016)问题。图 9 给出了多模态生成的示例。

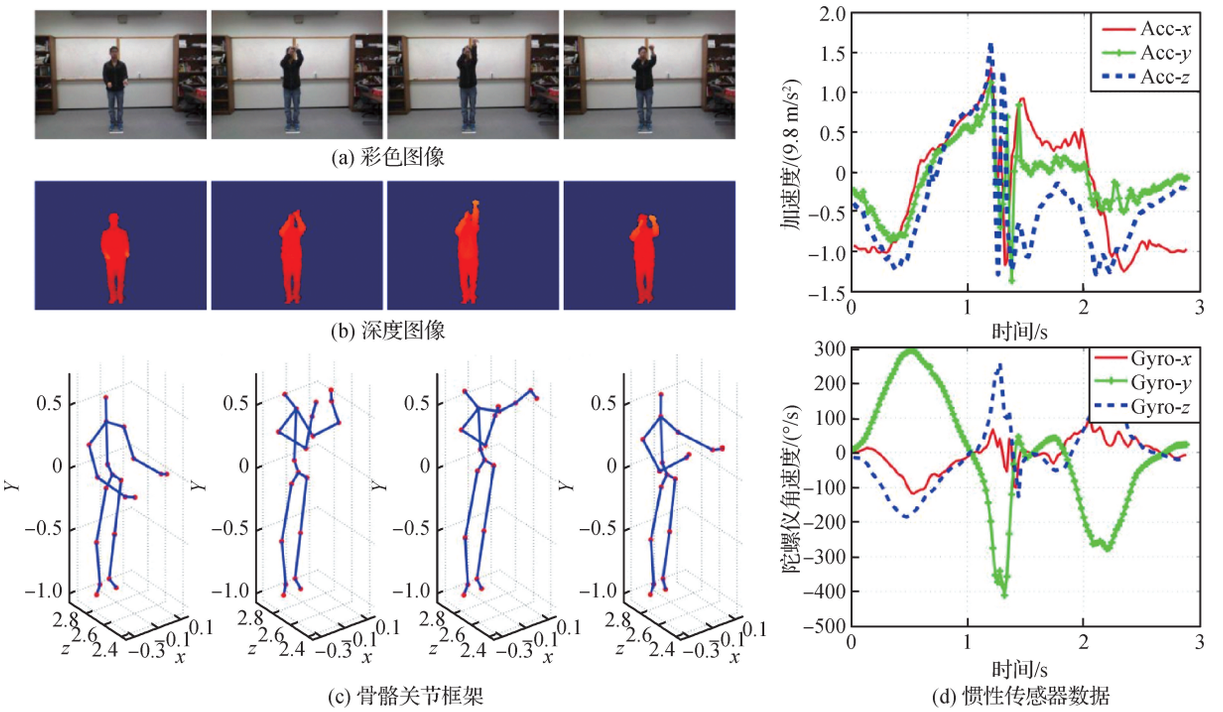


图 8　对应于投篮动作的多模态数据的一个例子(Chen 等,2015)

Fig. 8　An example of the multimodality data corresponding to the action basketball-shoot(Chen et al., 2015)

((a) the color images; (b) the depth images; (c) the skeleton joint frames; (d) the inertial sensor data)

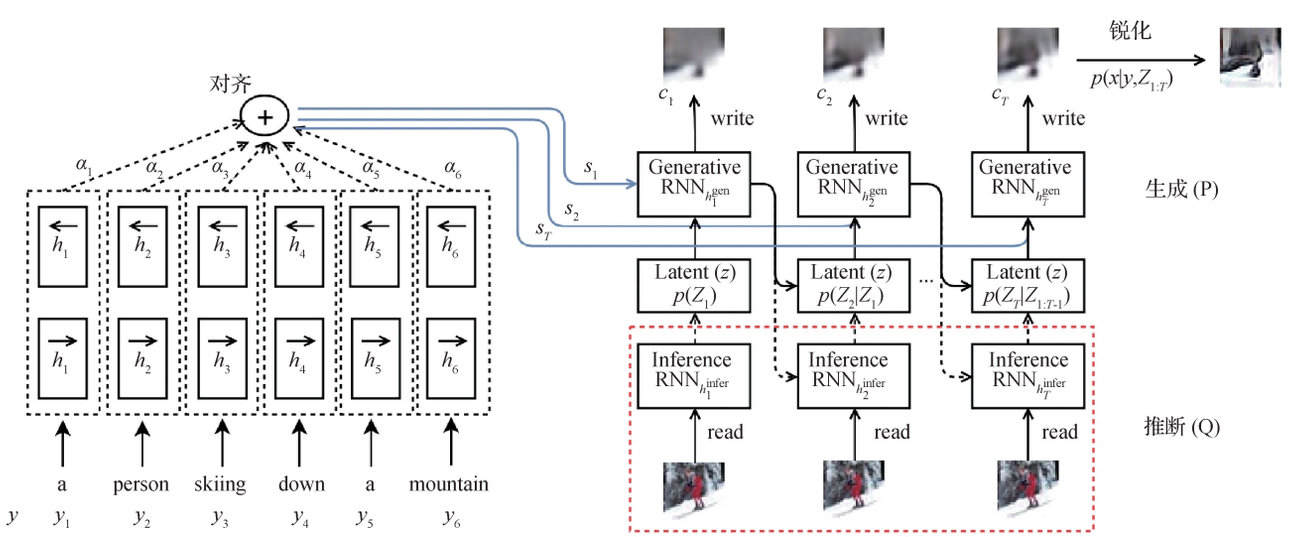


图 9　AlignDRAW 模型(Mansimov 等,2016)

Fig. 9　Alignment deep recurrent attention writer(AlignDRAW) model (Mansimov et al., 2016)

5)多媒体语言与视觉。还有一类多模态应用强调语言和视觉之间的相互作用。最具代表性的应用是视频中的语句时域定位(Gao 等,2017;Hendricks 等,2017;Yuan 等,2019b)、图像/视频标注(Duan 等,2018;Pan 等,2017b;You 等,2016)以及从自然语言(Liu 等,2019d;Mansimov 等,2016;Pan 等,2017a;Qiao 等,2019)生成图像/视频。语句时域定位是视频中动作探测的另一种形式,旨在利用自然语言描述而不是预定义的动作标签列表来识别视频中的特定活动(Gao 等,2017;Hendricks 等,2017;Yuan 等,2019b),因为复杂的人类活动不能简单地概括为有限的标签集。由于自然语言能够提供

对目标活动的更详细的描述,因此充分利用视觉和文本信号可以帮助更精确地检测时间边界(Chen 等,2018a;Zhang 等,2019)。这可以进一步促进一系列下游视频应用发展,如视频亮点检测(Badamdorj 等,2021;Yao 等,2016)、视频摘要(Narasimhan 等,2021;Yale 等,2015;Yuan 等,2019a)和视觉语言导航(Anderson 等,2018;Pashevich 等,2021)。此外,在图像区域中定位自然语言的概念类似于描述目标定位(Liu 等,2019a;Zeng 等,2020;Zhang 等,2018)。图像/视频标注旨在为输入图像/视频生成文本描述,其动机是帮助视力受损的人日常生活(Bigham 等,2010),并且对于基于内容的检索也非常重要。因此,标注技术可以应用于许多领域,包括生物医学、商业、军事、教育、数字图书馆和网络搜索(Hossain 等,2019)。反向任务上也取得了一些进展——从自然语言(Pan 等,2017a;Qiao 等,2019;Zhang 等,2017)生成图像/视频,其目的是提供更多渠道来增强媒体多样性。但是,图像/视频标注和生成任务在评估方面都存在主要挑战,即如何评估预测得到的描述结果或生成的图像/视频的质量。图 10 展示了视频标注的一个示例。

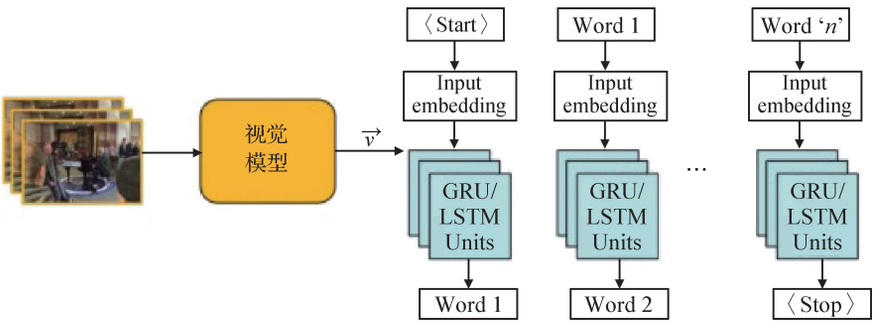


图 10　基于深度学习的视频标注的基本框架

Fig. 10　Framework of video annotation based on deep learning

2　机器学习推动多媒体发展

一方面,探索计算机算法类似人类的感知和推理能力一直是机器学习研究的首要任务之一。另一方面,人类的认知过程也可以视为感知和推理的级联(Manhaeve 等,2018;Peng 等,2017)。

1)人类探索周围环境,建立对世界的基本感知理解。

2)人类用人类所学到的知识来进一步推理人类的感知理解,并获得更深入的理解或新知识。

图 11 展示的是机器的类人认知过程,整个环节与上述人类认知过程极为相似。机器从任务环境和场景中获取原始数据,一边自下而上进行数据感知和特征提取,一边结合已有的知识对获得的特征进行符号化和符号推理,建立起对多媒体数据的进一步理解和认知。从思路上,它模仿了人类的认知过程,而人确实通过这一过程获取了认知能力;从技术上,它充分利用了外部环境的反馈和已有知识的信息,约束了整个学习过程从而能够达到收敛。

因此,一部分机器学习致力于研究感知和推理,以提高模型在特定任务和场景下的表现。尽管部分基于感知和推理的机器学习方法在提出时并不针对于多媒体数据和多媒体任务,但其中所蕴含的思路和策略可以迁移到多媒体领域中。这些相关技术的利用能够增强多媒体中类似人类的推理特征,从而产生推理能力更强的多媒体技术。

目前,深度学习方法已经可以很好地完成感知部分:可以区分猫和狗(Deng 等,2009),识别人类(Zhao 等,2017),并回答简单的问题(Antol 等,2015)。然而,它们几乎无法进行任何推理:既不能对其感知预测给出合理的解释,也不能进行明确的、人类可理解的推理。计算机算法离真正的类人感知和推理还很远,本节简要回顾了深度学习领域神经推理的进展,希望能为读者提供该方向的一些研究进展。

2.1　推理启发的感知学习

一些研究人员试图通过使用推理启发层或模块化来增强神经网络,从而使神经网络具备推理能力。例如,人类的推理过程可能包括多轮思考:人类可能

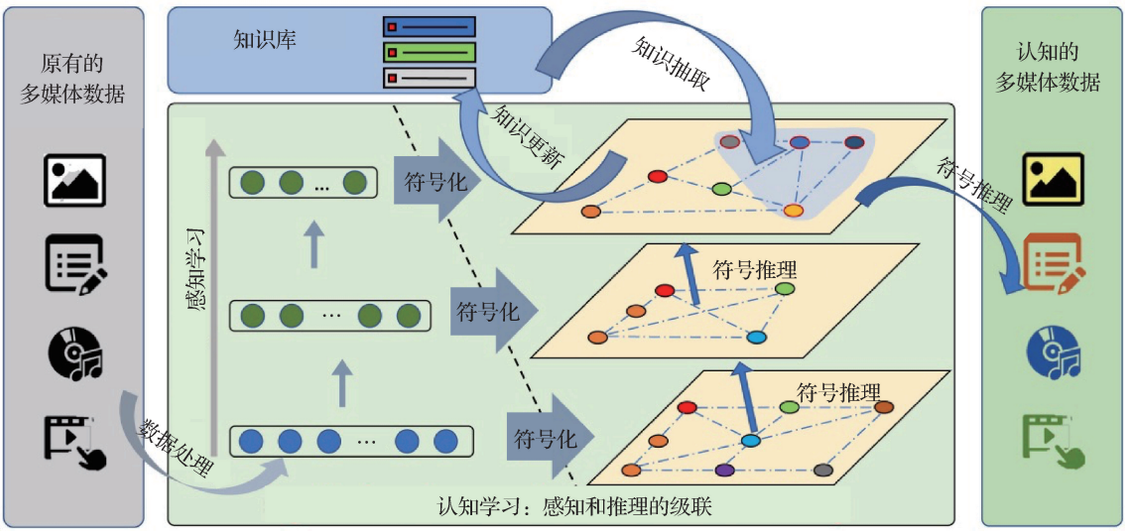


图 11　类人认知过程(Zhu 等,2020a)

Fig. 11　Human-like cognition(Zhu et al. , 2020a)

重复某个推理过程数次,直到达到某个目标。在这种情况下,可以将一些递归推理层添加到神经网络模型中,来模拟这一多轮过程。同时,关系信息和外部知识(以知识图谱的形式表示)对于机器学习算法获得对某些事实的推理能力也是必不可少的。在采用图神经网络(Scarselli 等,2009)或关系网络(Palm 等,2018;Santoro 等,2017)设计深度神经网络时,也考虑了这些因素。

1)多步推理。多步推理的目的是模仿人类的多步思维过程,可以利用递归神经网络(recursive neural network,RNN)实现,在这些工作中搜索者将一个递归单元作为多步推理模块插入神经网络(Cadene 等,2019;Hudson 和 Manning,2018;Wu 等,2018)。Hudson 等人(2018)设计了一个功能强大且复杂的递归单元,该单元能够满足递归神经网络单元的定义,并利用了许多直观的设计,如“控制单元”、“读取单元”和“写入单元”来模拟人的一步推理过程。Wu 等人(2018)采用多步推理策略来发现视觉问答(visual question answering,VQA)中分步推理的线索。Cadene 等人(2019)介绍了一种多步的多模态融合模式来回答 VQA 问题。此外,Das 等人(2019)提出了使用多步检索—读者交互模型来解决回答问题任务的方法。Duan 等人(2019a)使用多轮解码策略来学习更好的视频示例的项目特征。这些模型显著提高了对应性能,并自称是解决相关场景中问题的最新技术成果。然而,这些模型并不完美,因为它们需要更复杂的结构,而其内部推理过程也就更难解释。此外,这些方法采用固定的递归推理步骤以便于实现,这比人工推理过程要不灵活得多。图 12 展示了一个多步推理的例子。

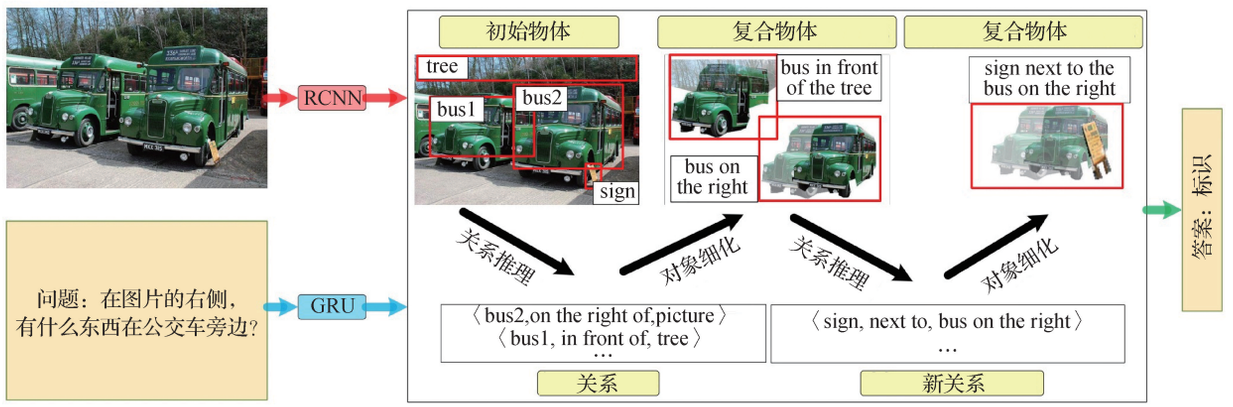


图 12　视觉问答的多步推理模型流程(Wu 等,2018)

Fig. 12　Chain of reasoning for VQA(Wu et al. , 2018)

2)关系推理。除了模仿人类的多步推理过程，模拟人类推理的另一种方式是利用图神经网络(graph neural networks，GNN)(Scarselli 等，2009)来模仿人类的关系推理能力。这些工作大多使用图神经网络聚合低级感知特征以得到增强的特征，来改进目标检测、目标跟踪和视觉问答任务(Chen 等，2019；Narasimhan 等，2018；Xiong 等，2019；Xu 等，2018；Xu，2019；Yu 等，2019b)的效果。Yu 等人(2019b)和 Xu 等人(2018)使用 GNN 来为各种任务集成目标检测方法中的特征。而 Narasimhan 等人(2018)和 Xiong 等人(2019)利用 GNN 作为一种消息传递工具来增强视觉问答对象的特征。除了图像特征方面的工作外，Liu 等人(2019c)和 Tsai 等人(2019)还基于时间和空间数据构建了用于基于视频的社会关系检测的图。Duan 等人(2019b)使用关系数据改进了 3 维点云分类任务的性能，并提高了模型的可解释性。在 Duan 等人(2019b)的工作中，一个对象可以看做是几个子对象的组合，这些子对象及其关系共同定义了该对象。例如，可以将鸟视为其子对象，如“翅膀”、“腿”、“头”、“身体”，及它们之间的关系的复杂集合，这被认为能够提高模型的性能和可解释性。此外，Wen 等人(2019)在多智能体强化学习任务中考虑了多智能体之间的关系，而 Chen 等人(2018b)则提出了一种将基于卷积和基于图的模型结合在一起的双流网络。图 13 展示了一个关系推理的例子。

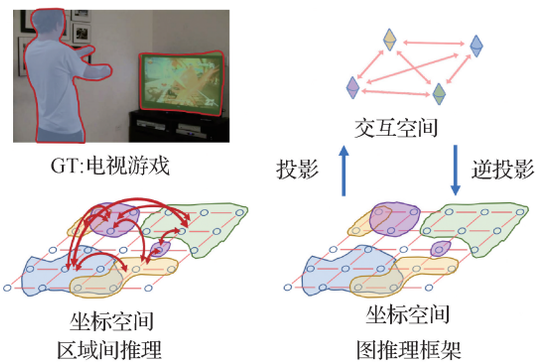


图 13　使用 GNN 作为目标检测任务中的推理工具(Chen 等，2019)

Fig. 13　GNNs as inference tools in object detection tasks (Chen et al., 2019)

3)注意力图和可视化。许多作品将注意力图作为推理的可视化或解释的一种方式，这些注意力图在一定程度上验证了相应方法的推理能力。特别是 Mascharka 等人(2018)提出将注意力图作为可视化和推理的线索。Cao 等人(2018)使用依赖树来指导 VQA 任务中注意力图的使用。Fan 和 Zhou(2018)利用潜在注意力图来改进多模型推理任务的效果。

2.2　感知—推理级联学习

一方面，相当多的研究致力于将推理能力集成到深层神经网络(deep neural network，DNN)中；另一方面，其他研究试图将 DNN 强大但初级的表示能力进行解耦，并将不同层次的感知过程进行级联来模拟高级的、人类可理解的认知，以期实现真正的强人工智能。

1)神经模块网络。神经模块网络(neural modular network，NMN)最早由 Andreas 等人(2016a)提出，并进一步在视觉推理任务中得到应用。NMN 的主要思想是使用一组预定义的神经模块来动态地组装针对某个实例特定的计算图，从而为每个输入实例实现个性化的异构计算。这些神经模块在设计时被赋予了特定的功能，例如查找、关联和回答等，它们通常根据不同的输入实例动态地被组装成层次树结构。

NMN 的设计动机来自两个观察结果：(1)视觉推理本质上是合成的；(2)深度神经网络具有强大的表示能力。

NMN 的组合特性允许将视觉推理过程分解为几个可共享、可重用的基本功能模块。然后，可以使用深度神经网络将这些原始功能模块作为神经模块来有效地实现。将视觉能力建模为分层原语的优点是多方面的。1)可以区分低级视觉感知和高级视觉推理；2)能够保持视觉世界的合成特性；3)与整体方法相比，这样生成的模型更具解释性，可能有利于未来人机回圈(human-in-the-loop，HITL)的多媒体智能的发展。

VQA 任务是开发计算机算法的视觉推理能力的一个很好的测试平台。广泛使用的 VQA 数据集(Antol 等，2015；Goyal 等，2017)更强调视觉感知而非视觉推理，这促使多个具有挑战性的多步、组合视觉推理数据集被设计出来(Hudson 和 Manning，2019；Johnson 等，2017a)。CLEVR(compositional language and elementary visual reasoning diagnostics dataset)数据集(Johnson 等，2017a)包括了一系列针

对合成图像的组合问题,这些合成图像仅由 3 类对象和 12 种不同属性(例如,蓝色大球体)所渲染,而 GQA 数据集(Hudson 和 Manning,2019)则处理具有更大语义空间和更多样化的视觉概念的真实图像。

最早由 Andreas 等人(2016a)提出的 NMN,利用异构、共同训练的神经模块组成深度网络。他们利用依赖关系解析器和手写规则生成模块布局,然后根据这些布局,使用一小部分模块组装成一个深度网络来回答视觉问题。动态模块网络(dynamic-neural modular network, D-NMN)(Andreas 等,2016b)是神经模块网络的后续工作,它学习了如何从手写规则所自动生成的一组候选布局中选择最佳布局。Hu 等人(2017)和 Johnson 等人(2017b)同时提出将布局预测问题转化为序列到序列的学习问题,而不是依赖现成的解析器来生成布局。这两种模型都可以预测网络布局,并都使用了强化和梯度下降相结合的方法来端到端地学习网络参数。值得注意的是,Johnson 等人(2017b)提出的模型为 CLEVR 数据集(Johnson 等,2017a)设计了细粒度的高度专业化的模块,例如 filter_rubber_material,它在模块实例化中硬编码文本参数。相比之下,Hu 等人(2017)提出的端到端模块网络(end-to-end module networks, N2 NMNs)模型设计了一组通用模块,例如查找、重定位,它们接受软注意力机制下的词嵌入作为文本参数。在 Hu 等人(2018)的后续工作——堆栈神经模块网络(stack neural modular network, Stack-NMN)中,网络不是对模块布局进行离散化的选择,而是使用完全可微的堆栈结构使布局光滑且连续。Mascharka 等人(2018)提出了设计透明网络(transparency by design network, TbD-net),该网络使用了与 Johnson 等人(2017b)类似的细粒度模块,但根据需求功能重新设计了每个模块。该模型不仅在 CLEVR 数据集(Johnson 等,2017a)上展现了近乎完美的性能,而且还展示了可以为模型行为提供解释的视觉注意力机制。

尽管这些模块化网络在合成图像上显示出近乎完美的准确性和可解释性,但在真实图像上执行全面的视觉推理仍然具有挑战性。Li 等人(2019)提出了感知视觉推理(perceptual visual reasoning, PVR)模型,用于对真实图像进行合成的、可解释的视觉推理,如图 14 所示。作者设计了一个包含从低级视觉感知到高级逻辑推理的丰富的通用模块库。同时,PVR 模型中的每个模块都能够从指导知识中感知外界监督,这有助于模块学习专门的和解耦的功能。他们在 GQA 数据集上的实验表明,PVR 模型可以在推理过程中产生透明、可解释的中间结果。

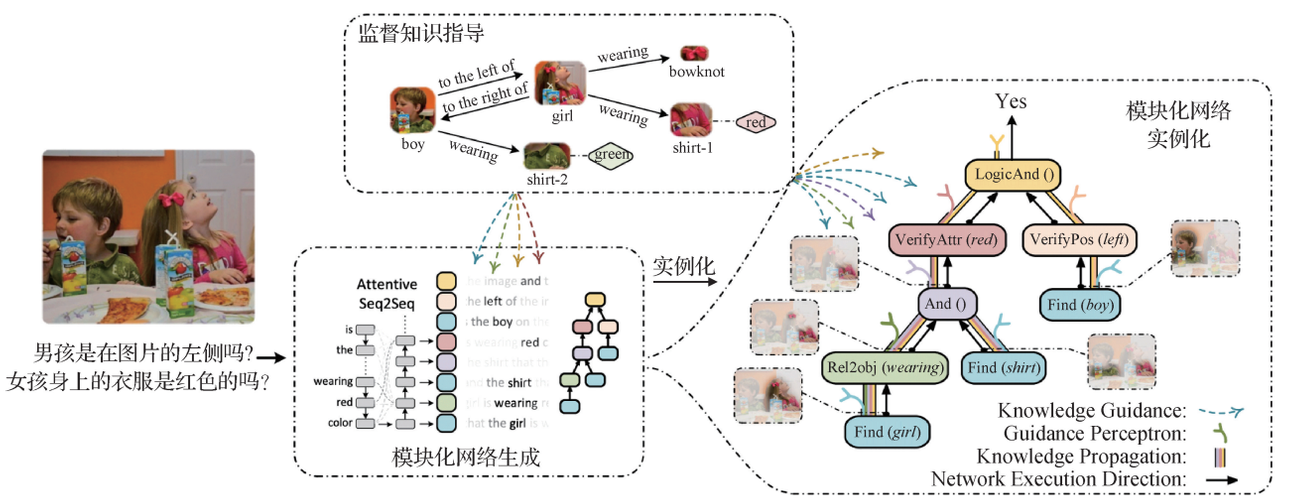


图 14 感知视觉推理(PVR)模型概述(Li 等,2019)

Fig. 14 Overview of perceptual visual reasoning (PVR) model(Li et al., 2019)

2)神经符号推理。除了组织按照语言学布局的模块化神经网络外,神经符号推理也是一个先进且有前途的方向,它的发展受到了来自于认知科学、人工智能、心理学以及结合了机器学习和自动推理的认知计算系统的推动。Garcez 等人(2002)引入了神经符号推理的基本思想:首先使用神经网络学习对场景的低级感知理解,然后将学习到的结果视为离散符号,在推理技术下进行

进一步推理。Yi 等人(2018)探索了视觉问答下的神经符号推理能力。视觉问答任务可以被解构为视觉概念探测、语言到程序的翻译和程序执行。通过学习视觉符号表征和语言符号表征，神经符号推理能够在预先设计的程序执行器中在视觉符号图上“执行”学习得到的语言符号代码来回答视觉问题。图 15 展示了神经符号推理的一个例子。

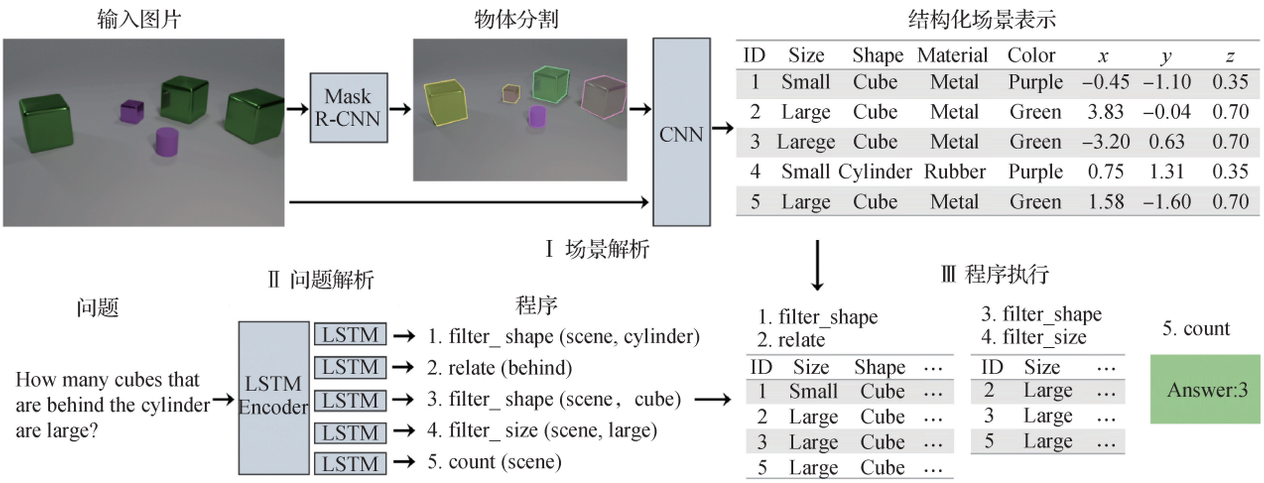


图 15　视觉问答的神经符号推理(Yi 等,2018)

Fig. 15　Neural symbolic reasoning for visual question answering(Yi et al., 2018)

神经符号推理因为其对 DNN 强大的特征表示能力和模拟人类高级推理和认知的能力的运用，在近年来受到了广泛的关注。然而，神经符号推理所使用的点对点(ad-hoc)程序设计和复杂的程序执行器严重限制了其性能，如何开发更好的程序设计和执行器值得进一步研究。

3　未来研究方向

3.1　多媒体图灵测试

本文介绍了多媒体智能的概念，并给出了多媒体和人工智能之间的循环(如图 1 所示)，在这个循环中，多媒体和人工智能相互影响。近年有诸多工作研究了从多媒体到 AI(机器学习)的半个循环，而从 AI(机器学习)到多媒体的另一半循环的研究却很少，从而显示出了循环的不完整性。本文认为多媒体图灵测试是能够补全循环的一种很有前途的方法。多媒体图灵测试包括视觉图灵测试(视觉和文本)(Geman 等,2015;Olague 等,2021;Qi 等,2015)、音频图灵测试(音频和文本)等，这些图灵测试在多种多媒体模态上进行。本节以视觉图灵测试为例，说明它与多媒体图灵测试中的其他成员是类似的。让计算机算法通过评估人类学习能力的视觉图灵测试可以作为进一步增强多媒体类人推理能力的一个阶段。视觉图灵测试的引入最初是在人类理解图像以及讲述图像故事的能力中得到的启发。在视觉图灵测试中，测试机器和人都被给予一幅图像和一系列问题，这些问题遵循着一条自然的故事线，这类似于人类在观看图像时所做出的反应。如果人类在测试中无法通过检查人和机器对给定图像的一系列问题的答案来区分人和机器，那么可以得出这样的结论：机器通过了视觉图灵测试。显然，通过视觉图灵测试需要类似人类的推理能力。

3.2　多媒体中的可解释推理

在今后的工作中，探索更具解释性的多媒体推理过程是一个值得进一步研究的重要方向。一种简单的方法是利用其他推理特征来扩充深度神经网络，从而丰富具有推理特性的深度神经网络。应该为深层神经网络配备更多更好的推理增强层或模块，这些模块将提高 DNN 的表示能力。例如，由于各种多媒体对象可以通过异构网络连接在一起，所以可以通过 GNN 来进行建模(Cai 等,2022a,b)。进一步，若能将 GNN 中的关系推理能力与类人多步推理相结合，也许可以开发出一种新的具有更强推理能力的 GNN 框架(Liu 等,2021;Zhao 等,2020)。而在更深层次上，类人认知学习(图 11 中的感知推理一体化学习)中最吸引人的是推理过程是透明和可解释的，这意味着人类知道模型如何以及为什么

会对某个场景起作用。因此,如何借助一阶逻辑、逻辑编程语言以至领域专用语言和更灵活的推理技术来设计更强大的推理模型值得进一步研究。此外,程序语言设计和程序执行器的自动化可以使神经符号推理在更复杂的场景中得到应用,这是在多媒体中实现可解释推理的另一种很有前景的方式(Trivedi等,2021;Verma 等,2018)。最后,鉴于当前的神经网络和推理模块是分开优化的,通过一个整体优化框架将神经网络和推理结合起来对于实现多媒体中可解释推理的目标起着重要作用。

3.3 自动机器学习与元学习

自动机器学习(automatic machine learning,AutoML)和元学习是学术界和工业界研究领域快速发展并受到大量关注的研究方向。AutoML 的目标是将端到端的机器学习模型在实际应用时进行自动化选取最优超参数(Akiba 等,2019;Bergstra 等,2011;Bergstra 和 Bengio,2012;Snoek 等,2012)与模型架构(Guo 等,2020b;Liu 等,2019b;Pham 等,2018;Zoph 和 Le,2017),从而使计算机算法能够自动适应不同的数据、任务和环境。AutoML 作为通用的机器学习方法,在多媒体领域中也大有可为,在多媒体数据表征、对齐、融合和迁移的过程中,存在着大量的超参数设置和模型架构选取,即使研究者有充分的专业经验和知识背景也可能花费大量时间和精力才能选出其中较为优秀的参数和架构。尤其当多媒体数据来源于真实的动态开放环境时,数据的分布将难以预先确定,因而更加需要模型自动应对当前的多媒体环境,实现感知和推理。

元学习(meta learning),即学会如何学习,旨在从不同的任务中提取和学习某种形式的一般知识,以供将来的各种其他任务使用,这也是人类特有的特征。关于元学习的现有文献主要集中在估计不同数据或任务之间的相似性,并试图借助额外的存储来尽可能多地记住以前的知识(Finn 等,2017;Rusu 等,2019;Santoro 等,2016)。在多媒体领域中的元学习方法能够增强多媒体模型实现迁移的能力,从而充分地将来自外部或内部的丰富感知结果向资源贫乏的模态进行迁移,提高相关模态的推理能力。

综上,将自动机器学习和元学习的思想应用于多模态多媒体问题,培养在类人任务和环境中的适应能力和知识迁移能力,是推动多媒体智能发展的另一个关键研究方向。

3.4 数字视网膜

在人类认知过程中,感知和推理之间实际上没有严格的界限——人类可能同时感知和推理(如图 11 所示)。因此,开发一些模拟这一过程的原型系统可能会将多媒体智能“循环”的补全推进一大步。

以现实世界的视频监控系统为例,当前系统中的视频流首先在摄像机上被采集并压缩,然后传输到后端服务器或云上进行大数据分析和检索。然而,人们认识到,压缩将不可避免地影响视觉特征的提取,从而降低后续分析和检索的性能。更重要的是,为了进行大数据分析和检索将来自数十万台摄像机的所有视频流聚合在一起是不切实际的。此时,类似人类的认知学习,即感知和推理的一体化,可以作为一种可能的解决方案,借鉴人类视网膜同时具有影像编码能力和特征编码能力的生物特征,人类能够设计一种更高效的摄像机,称为数字视网膜摄像机,简称数字视网膜(Gao 等,2021;Lou 等,2019)。这种新的数字视网膜的灵感来源于这样一个事实:生物视网膜实际上同时编码像素和特征,而大脑的下游区域接收的不是图像的一般像素表示,而是一组经过高度处理提取得到的特征。在数字视网膜框架下,相机通常配备全球统一的定时器和精确定位器,并且可以同时输出两个流,包括用于在线/离线观看和数据存储的压缩视频流,以及从原始图像/视频信号中提取的用于模式识别、视觉分析和搜索的压缩特征流。要实现数字视网膜需要有 3 项关键技术,包括分析友好的场景视频编码、视觉特征压缩描述符以及对视觉内容和特征进行整体压缩。通过只将特征流实时传输到云中心,这些摄像头能够为智能城市提供一个大规模的、像大脑一样的视觉系统(高文 等,2018;高文,2020a,b;Gao 等,2021;李赣湘,2021)。在多媒体的未来发展方向中,能够有效地在边缘设备进行感知和推理也是实现多媒体智能的重要因素(曹行健 等,2022;汪志涛,2022)。这一过程中包含了多媒体智能新的挑战和机遇,即在受限硬件设备下的数据建模和特征提取,这与寻常的多媒体数据表征在任务场景上是大相径庭的,但这一需求却是切实存在的。因此有必要将数字视网膜作为多媒体智能的未来发展方向之一,从而推动当前的多媒体研究朝着更实际的场景和更类人感知和推理的方向发展。

4 结 语

本文围绕“大数据”时代多媒体与人工智能融合的背景，提出了多媒体智能的新概念，探讨了多媒体和人工智能之间的相互影响，具体包括以下两个方向：

1）多媒体推动人工智能向着更具可解释性的方向发展；

2）人工智能促进多媒体推理能力的发展。

这两个方向形成了一个多媒体智能循环，其中多媒体和 AI 以交互和迭代的方式相互促进增强。本文讨论了每一循环中的研究进展，特别是研究多媒体如何推动机器学习发展以及机器学习如何反过来推动多媒体发展。最后，总结了循环中已经完成的工作，并指出完成循环未来所需要做的工作，然后对值得进一步深入探索的多媒体智能相关研究方向进行了思考。

参考文献（References）

Afouras T, Chung J S, Senior A, Vinyals O and Zisserman A. 2018. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence: #2889052 [DOI: 10.1109/TPAMI.2018.2889052]

Agrawal R, Faloutsos C and Swami A. 1993. Efficient similarity search in sequence databases//Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms. Chicago, USA: Springer: 69-84 [DOI: 10.1007/3-540-57301-1_5]

Akiba T, Sano S, Yanase T, Ohta T and Koyama M. 2019. Optuna: a next-generation hyperparameter optimization framework//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, USA: Association for Computing Machinery: 2623-2631 [DOI: 10.1145/3292500.3330701]

Anderson P, Wu Q, Teney D, Bruce J, Johnson M, Sünderhauf N, Reid I, Gould S and van den Hengel A. 2018. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 3674-3683 [DOI: 10.1109/CVPR.2018.00387]

Andreas J, Rohrbach M, Darrell T and Klein D. 2016a. Neural module networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 39-48 [DOI: 10.1109/CVPR.2016.12]

Andreas J, Rohrbach M, Darrell T and Klein D. 2016b. Learning to compose neural networks for question answering//Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics: 1545-1554 [DOI: 10.18653/v1/n16-1181]

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 (ICCV). Santiago, Chile: IEEE: 2425-2433 [DOI: 10.1109/ICCV.2015.279]

Badamdorj T, Rochan M, Wang Y and Cheng L. 2021. Joint visual and audio learning for video highlight detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 8107-8117 [DOI: 10.1109/ICCV48922.2021.00802]

Baevski A, Hsu W N, Conneau A and Auli M. 2021. Unsupervised speech recognition//Proceedings of the 35th Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 27826-27839

Baevski A, Zhou Y, Mohamed A and Auli M. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: [s.n.]

Bahdanau D, Cho K and Bengio Y. 2015. Neural machine translation by jointly learning to align and translate//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: [s.n.]

Baltrušaitis T, Ahuja C and Morency L P. 2019. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443 [DOI: 10.1109/TPAMI.2018.2798607]

Baroni M. 2015. Grounding distributional semantics in the visual world. Language and Linguistics Compass, 10(1): 3-13 [DOI: 10.1111/lnc3.12170]

Bergstra J, Bardenet R, Bengio Y and Kégl B. 2011. Algorithms for hyper-parameter optimization//Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada, Spain: Curran Associates Inc.: 2546-2554

Bergstra J and Bengio Y. 2012. Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13: 281-305

Bigham J P, Jayant C, Ji H J, Little G, Miller A, Miller R C, Miller R, Tatarowicz A, White B, White S and Yeh T. 2010. VizWiz: nearly real-time answers to visual questions//The 23rd Annual ACM Symposium on User Interface Software and Technology. New York, USA: Association for Computing Machinery: 333-342 [DOI: 10.1145/1866029.1866080]

Bojanowski P, Lajugie R, Grave E, Bach F, Laptev I, Ponce J and Schmid C. 2015. Weakly-supervised alignment of video with text//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 4462-4470 [DOI: 10.1109/ICCV.2015.507]

Cadene R, Ben-younes H, Cord M and Thome N. 2019. MUREL: multimodal relational reasoning for visual question answering//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 1989-1998 [DOI: 10.1109/CVPR.2019.00209]

Cai D S, Qian S S, Fang Q, Hu J, Ding W K and Xu C S. 2022a. Heterogeneous graph contrastive learning network for personalized micro-video recommendation. IEEE Transactions on Multimedia: #3151026 [DOI: 10.1109/TMM.2022.3151026]

Cai D S, Qian S S, Fang Q and Xu C S. 2022b. Heterogeneous hierarchical feature aggregation network for personalized micro-video recommendation. IEEE Transactions on Multimedia, 24: 805-818 [DOI: 10.1109/TMM.2021.3059508]

Cao Q X, Liang X D, Li B L, Li G B and Lin L. 2018. Visual question reasoning on general dependency tree//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 7249-7257 [DOI: 10.1109/CVPR.2018.00757]

Cao X J, Zhang Z T, Sun Y Z, Wang P, Xu S G, Liu F Q, Wang C, Peng F, Mu S Y, Liu W Y and Yang Y. 2022. The review of image processing and edge computing for intelligent transportation system. Journal of Image and Graphics, 27(6): 1743-1767 (曹行健, 张志涛, 孙彦赞, 王平, 徐树公, 刘富强, 王超, 彭飞, 穆世义, 刘文予, 杨铀. 2022. 面向智慧交通的图像处理与边缘计算. 中国图象图形学报, 27(6): 1743-1767) [DOI: 10.11834/jig.211266]

Chan W, Jaitly N, Le Q and Vinyals O. 2016. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE: 4960-4964 [DOI: 10.1109/ICASSP.2016.7472621]

Chen C, Jafari R and Kehtarnavaz N. 2015. UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor//Proceedings of 2015 IEEE International Conference on Image Processing (ICIP). Quebec City, Canada: IEEE: 168-172 [DOI: 10.1109/ICIP.2015.7350781]

Chen J Y, Chen X P, Ma L, Jie Z Q and Chua T S. 2018a. Temporally grounding natural sentence in video//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics: 162-171 [DOI: 10.18653/v1/d18-1015]

Chen L J, Lin S Y, Xie Y S, Lin Y Y and Xie X H. 2021. MVHM: a large-scale multi-view hand mesh benchmark for accurate 3D hand pose estimation//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 836-845 [DOI: 10.1109/WACV48630.2021.00088]

Chen X, Li L J, Li F F and Gupta A. 2018b. Iterative visual reasoning beyond convolutions//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 7239-7248 [DOI: 10.1109/CVPR.2018.00756]

Chen Y P, Rohrbach M, Yan Z C, Yan S C, Feng J S and Kalantidis Y. 2019. Graph-based global reasoning networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 433-442 [DOI: 10.1109/CVPR.2019.00052]

Ciaccia P, Patella M and Zezula P. 1997. M-tree: an efficient access method for similarity search in metric spaces//Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB). Athens, Greece: Morgan Kaufmann: 426-435

Cord M and Cunningham P. 2008. Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Berlin, Heidelberg, Germany: Springer [DOI: 10.1007/978-3-540-75171-7]

Das R, Dhuliawala S, Zaheer M and McCallum A. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Deng J, Dong W, Socher R, Li L J, Kai L 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 [DOI: 10.1109/CVPR.2009.5206848]

Ding C X and Tao D C. 2015. Robust face recognition via multimodal deep face representation. IEEE Transactions on Multimedia, 17(11): 2049-2058 [DOI: 10.1109/TMM.2015.2477042]

Duan X G, Huang W B, Gan C, Wang J D, Zhu W W and Huang J Z. 2018. Weakly supervised dense event captioning in videos//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 3063-3073

Duan X G, Wu Q, Gan C, Zhang Y W, Huang W B, van den Hengel A and Zhu W W. 2019a. Watch, reason and code: learning to represent videos using program//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 1543-1551 [DOI: 10.1145/3343031.3351094]

Duan Y Q, Zheng Y, Lu J W, Zhou J and Tian Q. 2019b. Structural relational reasoning of point clouds//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 949-958 [DOI: 10.1109/CVPR.2019.00104]

Ephrat A, Mosseri I, Lang O, Dekel T, Wilson K, Hassidim A, Freeman W T and Rubinstein M. 2018. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics, 37(4): #112 [DOI: 10.1145/3197517.3201357]

Escalera S, Baró X, Gonzàlez J, Bautista M A, Madadi M, Reyes M, Ponce-López V, Escalante H J, Shotton J and Guyon I. 2015. Chalearn looking at people challenge 2014: dataset and results//

Computer Vision – ECCV 2014 Workshops. Zurich, Switzerland: Springer: 459-473 [DOI: 10.1007/978-3-319-16178-5_32]

Fan H Q and Zhou J T. 2018. Stacked latent attention for multimodal reasoning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 1072-1080 [DOI: 10.1109/CVPR.2018.00118]

Finn C, Abbeel P and Levine S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR: 1126-1135

Frome A, Corrado G S, Shlens J, Bengio S, Dean J, Ranzato M A and Mikolov T. 2013. DeViSE: a deep visual-semantic embedding model//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 2121-2129

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 J Y, Sun C, Yang Z H and Nevatia R. 2017. TALL: temporal activity localization via language query//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5277-5285 [DOI: 10.1109/ICCV.2017.563]

Gao W. 2020a. City brain: challenges and solution. CAAI Transactions on Intelligent Systems, 15(4): 818-824 (高文. 2020a. 城市大脑的痛点与对策. 智能系统学报, 15(4): 818-824) [DOI: 10.11992/tis.202011038]

Gao W. 2020b. Digital retina, let smart city evolve from "see" to "understand". Scientific Chinese, (12): 30-31 (高文. 2020b. 数字视网膜, 让智慧城市从"看清"向"看懂"进化. 科学中国人, (12): 30-31)

Gao W, Ma S W, Duan L Y, Tian Y H, Xing P Y, Wang Y W, Wang S S, Jia H Z and Huang T J. 2021. Digital retina: a way to make the city brain more efficient by visual coding. IEEE Transactions on Circuits and Systems for Video Technology, 31(11): 4147-4161 [DOI: 10.1109/TCSVT.2021.3104305]

Gao W, Tian Y H and Wang J. 2018. Digital retina: revolutionizing camera systems for the smart city. Scientia Sinica Informationis, 48(8): 1076-1082 (高文, 田永鸿, 王坚. 2018. 数字视网膜: 智慧城市系统演进的关键环节. 中国科学: 信息科学, 48(8): 1076-1082) [DOI: 10.1360/N112018-00025]

Garcez A S D, Broda K B and Gabbay D M. 2002. Neural-Symbolic Learning Systems: Foundations and Applications. London, UK: Springer [DOI: 10.1007/978-1-4471-0211-3]

Garg A, Pavlovic V and Rehg J M. 2003. Boosted learning in dynamic Bayesian networks for multimodal speaker detection. Proceedings of the IEEE, 91(9): 1355-1369 [DOI: 10.1109/JPROC.2003.817119]

Geman D, Geman S, Hallonquist N and Younes L. 2015. Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences of the United States of America, 112(12): 3618-3623 [DOI: 10.1073/pnas.1422953112]

Ghahramani Z and Jordan M I. 1997. Factorial hidden markov models. Machine Learning, 29(2): 245-273 [DOI: 10.1023/A:1007425814087]

Gionis A, Indyk P and Motwani R. 1999. Similarity search in high dimensions via hashing//Proceedings of the 25th International Conference on Very Large Data Bases. Edinburgh, UK: Morgan Kaufmann Publishers Inc.: 518-529

Gönen M and Alpaydın E. 2011. Multiple kernel learning algorithms. The Journal of Machine Learning Research, 12: 2211-2268

Goyal Y, Khot T, Summers-Stay D, Batra D and Parikh D. 2017. Making the V in VQA matter: elevating the role of image understanding in visual question answering//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 6325-6334 [DOI: 10.1109/CVPR.2017.670]

Guo J Z, Zhu X Y, Zhao C X, Cao D, Lei Z and Li S Z. 2020a. Learning meta face recognition in unseen domains//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6162-6171 [DOI: 10.1109/CVPR42600.2020.00620]

Guo Z C, Zhang X Y, Mu H Y, Heng W, Liu Z C, Wei Y C and Sun J. 2020b. Single path one-shot neural architecture search with uniform sampling//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 544-560 [DOI: 10.1007/978-3-030-58517-4_32]

Gurban M, Thiran J P, Drugman T and Dutoit T. 2008. Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition//Proceedings of the 10th International Conference on Multimodal Interfaces. Chania, Greece: Association for Computing Machinery: 237-240 [DOI: 10.1145/1452392.1452442]

Hendricks L A, Wang O, Shechtman E, Sivic J, Darrell T and Russell B. 2017. Localizing moments in video with natural language//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5804-5813 [DOI: 10.1109/ICCV.2017.618]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Hossain Z, Sohel F, Shiratuddin M F and Laga H. 2019. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys, 51(6): #118 [DOI: 10.1145/3295748]

Hu R H, Andreas J, Darrell T and Saenko K. 2018. Explainable neural computation via stack neural module networks//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 55-71 [DOI: 10.1007/978-3-030-01234-2_4]

Hu R H, Andreas J, Rohrbach M, Darrell T and Saenko K. 2017. Learning to reason: end-to-end module networks for visual question

answering//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 804-813 [DOI: 10.1109/ICCV.2017.93]

Hudson D A and Manning C D. 2018. Compositional attention networks for machine reasoning//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Hudson D A and Manning C D. 2019. GQA: a new dataset for compositional question answering over real-world images [EB/OL]. [2022-01-12]. https://arxiv.org/pdf/1902.09506.pdf

Huo Y Q, Zhang M L, Liu G Z, Lu H Y, Gao Y Z, Yang G X, Wen J Y, Zhang H, Xu B G, Zheng W H, Xi Z Z, Yang Y Q, Hu A W, Zhao J M, Li R C, Zhao Y D, Zhang L, Song Y Q, Hong X, Cui W Q, Hou D Y, Li Y Y, Li J Y, Liu P Y, Gong Z, Jin C H, Sun Y C, Chen S Z, Lu Z W, Dou Z C, Jin Q, Lan Y Y, Zhao W X, Song R H and Wen J R. 2021. WenLan: bridging vision and language by large-scale multi-modal pre-training [EB/OL]. [2022-01-12]. https://arxiv.org/pdf/2103.06561.pdf

Johnson J, Hariharan B, van der Maaten L, Li F F, Zitnick C L and Girshick R. 2017a. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1988-1997 [DOI: 10.1109/CVPR.2017.215]

Johnson J, Hariharan B, van der Maaten L, Hoffman J, Li F F, Zitnick C L and Girshick R. 2017b. Inferring and executing programs for visual reasoning//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 3008-3017 [DOI: 10.1109/ICCV.2017.325]

Kahou S E, Bouthillier X, Lamblin P, Gulcehre C, Michalski V, Konda K, Jean S, Froumenty P, Dauphin Y, Boulanger-Lewandowski N, Chandias Ferrari R, Mirza M, Warde-Farley D, Courville A, Vincent P, Memisevic R, Pal C and Bengio Y. 2016. EmoNets: multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces, 10(2): 99-111 [DOI: 10.1007/s12193-015-0195-2]

Kalchbrenner N and Blunsom P. 2013. Recurrent continuous translation models//Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Washington, USA: Association for Computational Linguistics: 1700-1709

Khapra M M, Kumaran A and Bhattacharyya P. 2010. Everybody loves a rich cousin: an empirical study of transliteration through bridge languages//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA: Association for Computational Linguistics: 420-428

Kuznetsova A, Talati A, Luo Y W, Simmons K and Ferrari V. 2021. Efficient video annotation with visual interpolation and frame selection guidance//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3069-3078 [DOI: 10.1109/WACV48630.2021.00311]

Kuznetsova P, Ordonez V, Berg A C, Berg T L and Choi Y. 2012. Collective generation of natural image descriptions//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. Jeju Island, Korea(South): Association for Computational Linguistics: 359-368

Lafferty J D, McCallum A and Pereira F C N. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning. Williamstown, USA: Morgan Kaufmann Publishers Inc.: 282-289

Li B C, Wang Z, Liu J C and Zhu W W. 2013. Two decades of internet video streaming: a retrospective view. ACM Transactions on Multimedia Computing, Communications, and Applications, 9(1 s): #33 [DOI: 10.1145/2505805]

Li G H, Wang X and Zhu W W. 2019. Perceptual visual reasoning with knowledge propagation//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 530-538 [DOI: 10.1145/3343031.3350922]

Li G X. 2021. The application of "urban brain" in urban architectural planning. Urbanism and Architecture, 18(23): 79-81, 154 (李赣湘. 2021. 城市建筑规划中"城市大脑"的应用. 城市建筑, 18(23): 79-81, 154) [DOI: 10.19892/j.cnki.csjz.2021.23.21]

Li X J, Yin X, Li C Y, Zhang P C, Hu X W, Zhang L, Wang L J, Hu H D, Dong L, Wei F R, Choi Y and Gao J F. 2020. OSCAR: object-semantics aligned pre-training for vision-language tasks//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 121-137 [DOI: 10.1007/978-3-030-58577-8_8]

Lian D Z, Hu L N, Luo W X, Xu Y Y, Duan L X, Yu J Y and Gao S H. 2019. Multiview multitask gaze estimation with deep convolutional neural networks. IEEE Transactions on Neural Networks and Learning Systems, 30(10): 3010-3023 [DOI: 10.1109/TNNLS.2018.2865525]

Liu A A, Tian H S, Xu N, Nie W Z, Zhang Y D and Kankanhalli M. 2021. Toward region-aware attention learning for scene graph generation. IEEE Transactions on Neural Networks and Learning Systems: #3086066 [DOI: 10.1109/TNNLS.2021.3086066]

Liu D Q, Zhang H W, Zha Z J and Wang F L. 2019a. Referring expression grounding by marginalizing scene graph likelihood [EB/OL]. [2022-01-12]. https://arxiv.org/pdf/1906.03561.pdf

Liu H X, Simonyan K and Yang Y M. 2019b. DARTS: differentiable architecture search//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Liu X C, Liu W, Zhang M, Chen J W, Gao L L, Yan C G and Mei T. 2019c. Social relation recognition from videos via multi-scale spatial-temporal reasoning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 3561-3569 [DOI: 10.1109/CVPR.2019.

00368]

Liu Y, Wang X, Yuan Y T and Zhu W W. 2019d. Cross-modal dual learning for sentence-to-video generation//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 1239-1247 [DOI: 10.1145/3343031.3350986]

Lou Y H, Duan L Y, Luo Y, Chen Z Q, Liu T L, Wang S Q and Gao W. 2019. Towards digital retina in smart cities: a model generation, utilization and communication paradigm//Proceedings of 2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE: 19-24 [DOI: 10.1109/ICME.2019.00012]

Lu X, Zhu L, Liu L, Nie L Q and Zhang H X. 2021. Graph convolutional multi-modal hashing for flexible multimedia retrieval//Proceedings of the 29th ACM International Conference on Multimedia. [s. n.]: Association for Computing Machinery: 1414-1422 [DOI: 10.1145/3474085.3475598]

Ma L, Lu Z D, Shang L F and Li H. 2015. Multimodal convolutional neural networks for matching image and sentence//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 2623-2631 [DOI: 10.1109/ICCV.2015.301]

Malmaud J, Huang J, Rathod V, Johnston N, Rabinovich A and Murphy K. 2015. What's Cookin'? Interpreting cooking videos using text, speech and vision//Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, USA: Association for Computational Linguistics: 143-152 [DOI: 10.3115/v1/N15-1015]

Manhaeve R, Dumančić S, Kimmig A, Demeester T and De Raedt L. 2018. DeepProbLog: neural probabilistic logic programming//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 3753-3763

Mansimov E, Parisotto E, Ba L J and Salakhutdinov R. 2016. Generating images from captions with attention//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, USA: [s. n.]

Mascharka D, Tran P, Soklaski R and Majumdar A. 2018. Transparency by design: closing the gap between performance and interpretability in visual reasoning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 4942-4950 [DOI: 10.1109/CVPR.2018.00519]

McGurk H and MacDonald J. 1976. Hearing lips and seeing voices. Nature, 264(5588): 746-748 [DOI: 10.1038/264746a0]

Meng Q, Zhao S C, Huang Z and Zhou F. 2021. MagFace: a universal representation for face recognition and quality assessment//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 14220-14229 [DOI: 10.1109/CVPR46437.2021.01400]

Mikolov T, Sutskever I, Chen K, Corrado G and Dean J. 2013. Distributed representations of words and phrases and their compositionality//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 3111-3119

Mukherjee S S and Robertson N M. 2015. Deep head pose: gaze-direction estimation in multimodal video. IEEE Transactions on Multimedia, 17(11): 2094-2107 [DOI: 10.1109/TMM.2015.2482819]

Nagrani A, Yang S, Arnab A, Jansen A, Schmid C and Sun C. 2021. Attention bottlenecks for multimodal fusion//Proceedings of the 34th International Conference on Neural Information Processing Systems. [s. l.]: [s. n.]: 14200-14213

Nakov P and Ng H T. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages//Proceedings of 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, Singapore: Association for Computational Linguistics: 1358-1367

Narasimhan M, Lazebnik S and Schwing A G. 2018. Out of the box: reasoning with graph convolution nets for factual visual question answering//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 2659-2670

Narasimhan M, Rohrbach A and Darrell T. 2021. CLIP-it! language-guided video summarization//Proceedings of the 34th International Conference on Neural Information Processing Systems. [s. l.]: [s. n.]: 13988-14000

Natarajan P, Wu S, Vitaladevuni S, Zhuang X D, Tsakalidis S, Park U, Prasad R and Natarajan P. 2012. Multimodal feature fusion for robust event detection in web videos//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE: 1298-1305 [DOI: 10.1109/CVPR.2012.6247814]

Nefian A V, Liang L H, Pi X B, Liu X X, Mao C and Murphy K. 2002. A coupled HMM for audio-visual speech recognition//Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA: IEEE: 2013-2016 [DOI: 10.1109/ICASSP.2002.5745027]

Neverova N, Wolf C, Taylor G and Nebout F. 2016. ModDrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8): 1692-1706 [DOI: 10.1109/TPAMI.2015.2461544]

Ngiam J, Khosla A, Kim M, Nam J, Lee H and Ng A Y. 2011. Multimodal deep learning//Proceedings of the 28th International Conference on International Conference on Machine Learning. Washington, USA: Omnipress: 689-696

Ofli F, Chaudhry R, Kurillo G, Vidal R and Bajcsy R. 2013. Berkeley MHAD: a comprehensive multimodal human action database//Proceedings of 2013 IEEE Workshop on Applications of Computer Vision (WACV). Clearwater Beach, USA: IEEE: 53-60 [DOI: 10.1109/WACV.2013.6474999]

Olague G, Olague M, Jacobo-Lopez A R and Ibarra-Vázquez G. 2021. Less is more: pursuing the visual turing test with the kuleshov effect//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE: 1553-1561 [DOI: 10.1109/CVPRW53098.2021.00171]

Ordonez V, Kulkarni G and Berg T L. 2011. Im2Text: describing images using 1 million captioned photographs//Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada, Spain: Curran Associates Inc.: 1143-1151

Owens A, Isola P, McDermott J, Torralba A, Adelson E H and Freeman W T. 2016. Visually indicated sounds//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 2405-2413 [DOI: 10.1109/CVPR.2016.264]

Palm R B, Paquet U and Winther O. 2018. Recurrent relational networks//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 3372-3382

Pan Y W, Qiu Z F, Yao T, Li H Q and Mei T. 2017a. To create what you tell: generating videos from captions//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: Association for Computing Machinery: 1789-1798 [DOI: 10.1145/3123266.3127905]

Pan Y W, Yao T, Li H Q and Mei T. 2017b. Video captioning with transferred semantic attributes//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 984-992 [DOI: 10.1109/CVPR.2017.111]

Pashevich A, Schmid C and Sun C. 2021. Episodic transformer for vision-and-language navigation//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 15922-15932 [DOI: 10.1109/ICCV48922.2021.01564]

Peng Y X, Zhu W W, Zhao Y, Xu C S, Huang Q M, Lu H Q, Zheng Q H, Huang T J and Gao W. 2017. Cross-media analysis and reasoning: advances and directions. Frontiers of Information Technology and Electronic Engineering, 18(1): 44-57 [DOI: 10.1631/FITEE.1601787]

Petridis S, Stafylakis T, Ma P, Cai F P, Tzimiropoulos G and Pantic M. 2018. End-to-end audiovisual speech recognition//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 6548-6552 [DOI: 10.1109/ICASSP.2018.8461326]

Pham H, Guan M, Zoph B, Le Q and Dean J. 2018. Efficient neural architecture search via parameters sharing//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 4095-4104

Poria S, Chaturvedi I, Cambria E and Hussain A. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis//Proceedings of the 16th IEEE International Conference on Data Mining (ICDM). Barcelona, Spain: IEEE: 439-448 [DOI: 10.1109/ICDM.2016.0055]

Qi H, Wu T, Lee M W and Zhu S C. 2015. A restricted visual turing test for deep scene and event understanding [EB/OL]. [2022-01-12]. https://arxiv.org/pdf/1512.01715.pdf

Qiao T T, Zhang J, Xu D Q and Tao D C. 2019. MirrorGAN: learning text-to-image generation by redescription//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 1505-1514 [DOI: 10.1109/CVPR.2019.00160]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Rajendran J, Khapra M M, Chandar S and Ravindran B. 2016. Bridge correlational neural networks for multilingual multimodal representation learning//Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics: 171-181 [DOI: 10.18653/v1/N16-1021]

Ramachandram D and Taylor G W. 2017. Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Processing Magazine, 34(6): 96-108 [DOI: 10.1109/MSP.2017.2738401]

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2022-01-12]. https://arxiv.org/pdf/2204.06125.pdf

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8821-8831

Rasiwasia N, Moreno P J and Vasconcelos N. 2007. Bridging the gap: query by semantic example. IEEE Transactions on Multimedia, 9(5): 923-938 [DOI: 10.1109/TMM.2007.900138]

Rasiwasia N, Pereira J C, Coviello E, Doyle G, Lanckriet G R G, Levy R and Vasconcelos N. 2010. A new approach to cross-modal multimedia retrieval//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: Association for Computing Machinery: 251-260 [DOI: 10.1145/1873951.1873987]

Reed S, Akata Z, Yan X C, Logeswaran L, Schiele B and Lee H. 2016. Generative adversarial text to image synthesis//Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, USA: JMLR.org: 1060-1069

Rusu A A, Rao D, Sygnowski J, Vinyals O, Pascanu R, Osindero S and Hadsell R. 2019. Meta-learning with latent embedding optimization//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Santoro A, Bartunov S, Botvinick M M, Wierstra D and Lillicrap T P. 2016. Meta-learning with memory-augmented neural networks//Proceedings of the 33rd International Conference on Machine Learning. New York, USA: JMLR.org: 1842-1850

Santoro A, Raposo D, Barrett D G T, Malinowski M, Pascanu R, Battaglia P and Lillicrap T. 2017. A simple neural network module for relational reasoning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 4974-4983

Scarselli F, Gori M, Tsoi A C, Hagenbuchner M and Monfardini G. 2009. The graph neural network model. IEEE Transactions on Neural Networks, 20(1): 61-80 [DOI: 10.1109/TNN.2008.2005605]

Schultz P T and Sartini R A. 2016. Method and system for multi-factor biometric authentication. U.S., No. 9 323 912

Singh B, Marks T K, Jones M, Tuzel O and Shao M. 2016. A multi-stream Bi-directional recurrent neural network for fine-grained action detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 1961-1970 [DOI: 10.1109/CVPR.2016.216]

Sitová Z, Šeděnka J, Yang Q, Peng G, Zhou G, Gasti P and Balagani K S. 2016. HMOG: new behavioral biometric features for continuous authentication of smartphone users. IEEE Transactions on Information Forensics and Security, 11(5): 877-892 [DOI: 10.1109/TIFS.2015.2506542]

Snoek J, Larochelle H and Adams R P. 2012. Practical bayesian optimization of machine learning algorithms//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 2951-2959

Socher R, Ganjoo M, Manning C D and Ng A Y. 2013. Zero-shot learning through cross-modal transfer//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 935-943

Srivastava N and Salakhutdinov R. 2012. Learning representations for multimodal data with deep belief nets//Proceedings of 2012 International Conference on Machine Learning Workshop. Edinburgh, UK: [s.n.]: 978-971

Sutskever I, Vinyals O and Le Q V. 2014. Sequence to sequence learning with neural networks//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 3104-3112

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 (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics: 5100-5111 [DOI: 10.18653/v1/D19-1514]

Trivedi D, Zhang J, Sun S H and Lim J J. 2021. Learning to synthesize programs as interpretable and generalizable policies//Proceedings of the 34th International Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 25146-25163

Tsai Y H H, Divvala S, Morency L P, Salakhutdinov R and Farhadi A. 2019. Video relationship reasoning using gated spatio-temporal energy graph//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 10416-10425 [DOI: 10.1109/CVPR.2019.01067]

van den Oord A, Dieleman S and Schrauwen B. 2013. Deep content-based music recommendation//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 2643-2651

van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A W and Kavukcuoglu K. 2016. WaveNet: a generative model for raw audio//Proceedings of the 9th ISCA Speech Synthesis Workshop. Sunnyvale, USA: ISCA: #125

Venugopalan S, Xu H J, Donahue J, Rohrbach M, Mooney R and Saenko K. 2015. Translating videos to natural language using deep recurrent neural networks//Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA: Association for Computational Linguistics: 1494-1504 [DOI: 10.3115/v1/N15-1173]

Verma A, Murali V, Singh R, Kohli P and Chaudhuri S. 2018. Programmatically interpretable reinforcement learning//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 5045-5054

Vinyals O, Toshev A, Bengio S and Erhan D. 2015. Show and tell: a neural image caption generator//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 3156-3164 [DOI: 10.1109/CVPR.2015.7298935]

Wang D X, Cui P, Ou M D and Zhu W W. 2015. Deep multimodal hashing with orthogonal regularization//Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI Press: 2291-2297

Wang J D, Zhang T, Song J K, Sebe N and Shen H T. 2018. A survey on learning to hash. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4): 769-790 [DOI: 10.1109/TPAMI.2017.2699960]

Wang K Y, Yin Q Y, Wang W, Wu S and Wang L. 2016a. A comprehensive survey on cross-modal retrieval [EB/OL]. [2022-01-12]. https://arxiv.org/pdf/1607.06215.pdf

Wang X, Donaldson R, Nell C, Gorniak P, Ester M and Bu J J. 2016b. Recommending groups to users using user-group engagement and time-dependent matrix factorization//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI: 1331-1337

Wang X, Hoi S C H, Ester M, Bu J J and Chen C. 2017. Learning personalized preference of strong and weak ties for social recommendation//Proceedings of the 26th International Conference on World Wide Web. Perth, Australia: International World Wide Web Conferences Steering Committee: 1601-1610 [DOI: 10.1145/3038912.3052556]

Wang X, Lu W, Ester M, Wang C and Chen C. 2016c. Social recommendation with strong and weak ties//Proceedings of the 25th ACM

International on Conference on Information and Knowledge Management. Indianapolis, USA: Association for Computing Machinery: 5-14 [DOI: 10.1145/2983323.2983701]

Wang X, Zhu W W and Liu C H. 2019a. Social recommendation with optimal limited attention//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, USA: Association for Computing Machinery: 1518-1527 [DOI: 10.1145/3292500.3330939]

Wang X, Zhu W W and Liu C H. 2019b. Semi-supervised deep quantization for cross-modal search//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 1730-1739 [DOI: 10.1145/3343031.3350934]

Wang Y K, Huang W B, Sun F C, Xu T Y, Rong Y and Huang J Z. 2020. Deep multimodal fusion by channel exchanging//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #406

Wang Z T. 2022. Design and application of intelligent road traffic system based on multi-access edge computing. Traffic and Transportation, 38(3): 50-54 (汪志涛. 2022. 基于边缘计算的智能道路交通系统设计及应用. 交通与运输, 38(3): 50-54)

Wen Y, Yang Y D, Luo R, Wang J and Pan W. 2019. Probabilistic recursive reasoning for multi-agent reinforcement learning//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Wöllmer M, Kaiser M, Eyben F, Schuller B and Rigoll G. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, 31(2): 153-163 [DOI: 10.1016/j.imavis.2012.03.001]

Wu C F, Liu J L, Wang X J and Dong X. 2018. Chain of reasoning for visual question answering//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 273-283

Wu D, Pigou L, Kindermans P J, Le N D H, Shao L, Dambre J and Odobez J M. 2016. Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8): 1583-1597 [DOI: 10.1109/TPAMI.2016.2537340]

Xiong P X, Zhan H Y, Wang X, Sinha B and Wu Y. 2019. Visual query answering by entity-attribute graph matching and reasoning//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 8349-8358 [DOI: 10.1109/CVPR.2019.00855]

Xu H, Jiang C H, Liang X D, Lin L and Li Z G. 2019. Reasoning-RCNN: unifying adaptive global reasoning into large-scale object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 6412-6421 [DOI: 10.1109/CVPR.2019.00658]

Xu M M, Zhao C, Rojas D S, Thabet A and Ghanem B. 2020a. G-TAD: sub-graph localization for temporal action detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 10153-10162 [DOI: 10.1109/CVPR42600.2020.01017]

Xu M Z, Xiong Y J, Chen H, Li X Y, Xia W, Tu Z W and Soatto S. 2021. Long short-term transformer for online action detection//Proceedings of the 34th International Conference on Neural Information Processing Systems. [s.l.]: [s.n.]: 1086-1099

Xu Q T, Likhomanenko T, Kahn J, Hannun A Y, Synnaeve G and Collobert R. 2020b. Iterative pseudo-labeling for speech recognition//Proceedings of Interspeech 2020, 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 1006-1010 [DOI: 10.21437/Interspeech.2020-1800]

Xu Y L, Qin L, Liu X B, Xie J W and Zhu S C. 2018. A causal and-or graph model for visibility fluent reasoning in tracking interacting objects//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 2178-2187 [DOI: 10.1109/CVPR.2018.00232]

Yale S, Vallmitjana J, Stent A and Jaimes A. 2015. TVSum: summarizing web videos using titles//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 5179-5187 [DOI: 10.1109/CVPR.2015.7299154]

Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H and Courville A. 2015. Describing videos by exploiting temporal structure//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 4507-4515 [DOI: 10.1109/ICCV.2015.512]

Yao T, Mei T and Rui Y. 2016. Highlight detection with pairwise deep ranking for first-person video summarization//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 982-990 [DOI: 10.1109/CVPR.2016.112]

Yi K X, Wu J J, Gan C, Torralba A, Kohli P and Tenenbaum J B. 2018. Neural-symbolic VQA: disentangling reasoning from vision and language understanding//Proceedings of the 32nd International Conference Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 1039-1050

You Q Z, Jin H L, Wang Z W, Fang C and Luo J B. 2016. Image captioning with semantic attention//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 4651-4659 [DOI: 10.1109/CVPR.2016.503]

Yu H N, Wang J, Huang Z H, Yang Y and Xu W. 2016. Video paragraph captioning using hierarchical recurrent neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 4584-4593 [DOI: 10.1109/CVPR.2016.496]

Yu S Z, Wang X, Zhu W W, Cui P and Wang J D. 2019a. Disparity-preserved deep cross-platform association for cross-platform video recommendation//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI.org: 4635-

4641 [DOI: 10.24963/ijcai.2019/644]

Yu W J, Liang X D, Gong K, Jiang C H, Xiao N and Lin L. 2019b. Layout-graph reasoning for fashion landmark detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 2932-2940 [DOI: 10.1109/CVPR.2019.00305]

Yuan Y T, Mei T, Cui P and Zhu W W. 2019a. Video summarization by learning deep side semantic embedding. IEEE Transactions on Circuits and Systems for Video Technology, 29(1): 226-237 [DOI: 10.1109/TCSVT.2017.2771247]

Yuan Y T, Mei T and Zhu W W. 2019b. To find where you talk: temporal sentence localization in video with attention based location regression//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press: 9159-9166 [DOI: 10.1609/aaai.v33i01.33019159]

Yuhas B P, Goldstein M H and Sejnowski T J. 1989. Integration of acoustic and visual speech signals using neural networks. IEEE Communications Magazine, 27(11): 65-71 [DOI: 10.1109/35.41402]

Zeng R H, Xu H M, Huang W B, Chen P H, Tan M K and Gan C. 2020. Dense regression network for video grounding//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 10284-10293 [DOI: 10.1109/CVPR42600.2020.01030]

Zhang D, Dai X Y, Wang X, Wang Y F and Davis L S. 2019. MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 1247-1257 [DOI: 10.1109/CVPR.2019.00134]

Zhang H, Xu T, Li H S, Zhang S T, Wang X G, Huang X L and Metaxas D. 2017. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5908-5916 [DOI: 10.1109/ICCV.2017.629]

Zhang H W, Niu Y L and Chang S F. 2018. Grounding referring expressions in images by variational context//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4158-4166 [DOI: 10.1109/CVPR.2018.00437]

Zhang J H, Zhu Y Q, Liu Q, Wu S, Wang S H and Wang L. 2021. Mining latent structures for multimedia recommendation//Proceedings of the 29th ACM International Conference on Multimedia. [s.l.]: Association for Computing Machinery: 3872-3880 [DOI: 10.1145/3474085.3475259]

Zhang L and Rui Y. 2013. Image search—from thousands to billions in 20 years. ACM Transactions on Multimedia Computing, Communications, and Applications, 9(1s): #36 [DOI: 10.1145/2490823]

Zhang W, Zhang Y M, Ma L, Guan J W and Gong S J. 2015. Multimodal learning for facial expression recognition. Pattern Recognition, 48(10): 3191-3202 [DOI: 10.1016/j.patcog.2015.04.012]

Zhang X C, Park S, Beeler T, Bradley D, Tang S Y and Hilliges O. 2020. ETH-XGaze: a large scale dataset for gaze estimation under extreme head pose and gaze variation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 365-381 [DOI: 10.1007/978-3-030-58558-7_22]

Zhao J W, Han R Z, Gan Y Y, Wan L, Feng W and Wang S. 2020. Human identification and interaction detection in cross-view multi-person videos with wearable cameras//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: Association for Computing Machinery: 2608-2616 [DOI: 10.1145/3394171.3413903]

Zhao L M, Li X, Zhuang Y T and Wang J D. 2017. Deeply-learned part-aligned representations for person re-identification//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 3239-3248 [DOI: 10.1109/ICCV.2017.349]

Zhu W W, Cui P, Wang Z and Hua G. 2015. Multimedia big data computing. IEEE Multimedia, 22(3): #96 [DOI: 10.1109/MMUL.2015.66]

Zhu W W, Wang X and Gao W. 2020a. Multimedia intelligence: when multimedia meets artificial intelligence. IEEE Transactions on Multimedia, 22(7): 1823-1835 [DOI: 10.1109/TMM.2020.2969791]

Zhu W W, Wang X and Li H Z. 2020b. Multi-modal deep analysis for multimedia. IEEE Transactions on Circuits and Systems for Video Technology, 30(10): 3740-3764 [DOI: 10.1109/TCSVT.2019.2940647]

Zoph B and Le Q V. 2017. Neural architecture search with reinforcement learning//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview.net

作者简介

朱文武，1963 年生，男，教授，主要研究方向为多媒体大数据、多媒体智能。
E-mail：wwzhu@tsinghua.edu.cn

田永鸿，通信作者，男，教授，主要研究方向为分布式机器学习、神经形态视觉和视频大数据。
E-mail：yhtian@pku.edu.cn

王鑫，男，助理研究员，主要研究方向为多媒体智能分析、机器学习。E-mail：xin_wang@tsinghua.edu.cn

高文，男，中国工程院院士，主要研究方向为人工智能、模式识别与多媒体计算。E-mail：wgao@pku.edu