中国科学: 信息科学 2023 年 第 53 卷 第 1 期: 1–32
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

多模态认知计算

李学龙[1,2]

1. 西北工业大学光电与智能研究院, 西安 710072
2. 智能交互与应用工业和信息化部重点实验室(西北工业大学), 西安 710072
E-mail: li@nwpu.edu.cn

收稿日期: 2022–06–08; 接受日期: 2022–07–21; 网络出版日期: 2023–01–11

国家自然科学基金 (批准号: 61871470) 资助项目

摘要 人类利用视觉、听觉等多种感官理解周围环境, 通过整合多种感知模态, 形成对事件的整体认识. 为使机器更好地模仿人类的认知能力, 多模态认知计算模拟人类的 "联觉" (synaesthesia), 探索图像、视频、文本、语音等多模态输入的高效感知与综合理解手段, 是人工智能领域的重要研究内容, 也是实现 "通用人工智能" 的关键之一. 近年来, 随着多模态时空数据的海量爆发和计算能力的快速提升, 国内外学者提出了大量方法, 以应对日益增长的多样化需求. 然而, 当前的多模态认知计算仍局限于人类表观能力的模仿, 缺乏认知层面的理论依据. 本文从信息论角度出发, 建立了认知过程的信息传递模型, 结合信容 (information capacity), 提出了多模态认知计算能够提高机器的信息提取能力这一观点, 从理论上对多模态认知计算各项任务进行了统一. 进而, 根据机器对多模态信息的认知模式, 从多模态关联、跨模态生成和多模态协同这 3 个方面对现有方法进行了梳理与总结, 系统地分析了其中的关键问题与解决方案. 最后, 结合当前阶段人工智能的发展特点, 重点思考多模态认知计算领域面临的难点与挑战, 并对未来发展趋势进行了深入分析与展望.

关键词 人工智能, 多模态, 认知计算, 联觉, 信容

1 引言

让机器像人类一样智能地感知周围环境并做出决策, 是人工智能的目标之一. 在对信息的处理模式上, 人类与机器存在巨大差异. 为构建模拟人类认知模式的智能系统, 英国 Ulster 大学的研究者在 2003 年将 "认知计算" (cognitive computing) 的概念引入信息领域, 重点关注认知科学与传统的视音频、图像、文本等处理之间互相联系的机理和机制, 并且开设了相应的教学课程 [1]. 在 21 世纪初, Xuelong Li 创立了 IEEE-SMC 认知计算技术委员会, 当时为认知计算设定的目标是 "Cognitive Computing breaks the traditional boundary between neuroscience and computer science, and paves the way for machines that will have reasoning abilities analogous to a human brain. It's an interdisciplinary

引用格式: 李学龙. 多模态认知计算. 中国科学: 信息科学, 2023, 53: 1–32, doi: 10.1360/SSI-2022-0226
Li X L. Multi-modal cognitive computing (in Chinese). Sci Sin Inform, 2023, 53: 1–32, doi: 10.1360/SSI-2022-0226

© 2023《中国科学》杂志社

research and application field, and uses methods from psychology, biology, signal processing, physics, information theory, mathematics, and statistics. The development of Cognitive Computing will cross-fertilize these other research areas with which it interacts. There are many open problems to be addressed and to be defined. This technical committee tackles these problems in both academia and industry, and focuses on new foundations/technologies that are intrinsic to Cognitive Computing."[1)] 二十年来, 认知计算逐渐受到各领域学者的关注.

在现实生活中, 人类利用视觉、听觉、触觉等多种感官认识世界, 不同感官刺激交融形成统一的多感觉体验. 这种多感官协作对于机器而言即为 "多模态". 认知神经学研究 [2] 表明, 一类感官刺激可能会作用于其他感官通道, 这种现象被称为 "联觉" (synaesthesia). 2008 年, Li 等 [3] 在 "Visual music and musical vision" 一文中首次将联觉引入信息领域, 并从信息度量角度计算多模态数据的关联, 尝试性地探讨了 "多模态认知计算" 的理论及应用. 随着人工智能第三次发展高潮的影响逐渐深化, 多模态认知计算迎来了新的发展机遇, 成为航空航天、智能制造、医疗健康等重大领域共同关注的研究课题, 对推动我国人工智能战略发展具有重要意义. 在国内, 相应的研究和探索也有较长的历史和积累, 有很多顶尖的研究团队. 2008 年, 国家自然科学基金委员会设立的重大研究计划 "视听觉信息的认知计算", 实施以来取得了丰硕成果. 2017 年, 国务院印发了《新一代人工智能发展规划》, 明确提出 "建立大规模类脑智能计算的新模型和脑启发的认知计算模型", 研究 "以自然语言理解和图像图形为核心的认知计算理论和方法". 当前, 多模态认知计算研究已从学术牵引转化为需求牵引, 在图像、视频、文本、语音等海量多模态数据和强大算力的支撑下, 国内外各大知名企业与研究机构纷纷加入此项研究中. 然而, 在蓬勃发展的背后, 多模态认知计算的理论机理仍不明确. 认知神经学家提出了大量理论与假设来刻画人类对多感知模态的认知过程. 而在信息领域, 多模态认知计算仍停留在人类认知的观察和模仿阶段, 缺乏机理性解释与统一的学习理论框架.

本文尝试以认知为切入点, 阐释多模态认知计算的理论意义. 认知是人类从现实世界中提取并加工信息的过程, 外界信息通过视、听、嗅、味、触等多种感知通道传送到大脑, 对大脑皮层产生刺激. 神经科学相关研究 [4] 表明, 多种感官刺激的联合作用会产生 "整体大于局部之和" 的效果. 例如, 在观看影视剧时, 画面和声音的同时刺激会给人类带来深刻、全面的感受, 也帮助人类更准确地理解影视内容. 这种现象是如何产生的? 认知科学研究 [5] 指出, 人类在接收外界刺激时会选择性地关注其中的一部分. 这种 "注意力机制" 作为人类认知能力的重要组成部分, 有效提高了信息加工的效率. 当影视画面与声音同步时, 人类的注意力并不会被分散, 而会集中在影视剧中发生的事件上, 视觉与听觉感官同时得到了关注. 基于上述观察, 本文提出以下假设: 当同一事件引起多种感官的同步刺激时, 不同感官通道共享注意力, 人类可以感知更多信息. 从认知计算角度出发, 本文利用信息论的理论对上述假设进行建模. 信息论奠基人 C. Shannon 在 1948 年的文章 "A mathematical theory of communication" 中提出了信息熵的概念, 用其表示随机变量的不确定程度, 为信息量的度量提供了解决方案. 根据信息熵定义, 假设事件空间 X 的概率分布已知, 事件 x 的概率为 $p(x)$, 其所带来的信息量为

$$h(x) = -\log p(x). \tag{1}$$

事件的概率越小, 其发生所提供的信息量越大. 例如, 红色天空比蓝色天空出现的概率小, 一般来说其信息量也就相对更大. 同时, 在不同认知任务中, 事件的发生概率存在差异, 提供的信息量也有所区别. 例如, 红色天空为气象学研究带来的信息量要高于其对心理学研究提供的信息量. 对于给定认知任务

1) https://www.ieeesmc.org/technical-activities/human-machine-systems/cognitive-computing.

表 1 代表性多模态认知计算模型训练数据量

Table 1 Amount of training images of some representative multi-modal cognitive models

Method	Amount of training images	Task	Year
VMMV [3]	3542	Image-music retrieval	2008
SCM [8]	2173	Text-image retrieval	2010
GMMFA [9]	2808	Text-image retrieval	2012
DFE [10]	6092	Text-image retrieval	2014
AlignDRA [11]	82783	Text-image synthesis/retrieval	2016
I2S [12]	36646	Image-lyrics retrieval	2017
DRAU [13]	123287	Visual question answering	2019
M6 [6]	60500000	General	2021
CLIP [7]	400000000	General	2021

T, 事件 x 提供的信息量为

$$h(x,T) = -\log p(x|_T), \tag{2}$$

其中 $p(x|_T)$ 为 x 在给定任务 T 的情况下发生的概率. 式 (2) 与条件熵在形式上类似, 但在概念上有本质区别: 条件熵在给定变量情况下, 计算未知变量的信息量; 而式 (2) 计算的是事件对于特定任务的信息量. 假设事件空间 $X \in \mathbb{R}^{m\times s\times t}$ 为感知模态 (m), 空间 (s), 时间 (t) 上的张量, 受上述现象启发, 本文将个体从事件空间中获得的信息量定义为

$$K = \left\| A \odot \sum_{i=1}^{m} I_i \right\|, \tag{3}$$

其中矩阵 $I_i \in \mathbb{R}^{s\times t}$ 为第 i 个模态所有事件的信息量所构成的矩阵, 计算为 $I_{ijk} = h(X_{ijk}, T)$. $\odot$ 代表矩阵点乘运算, $A \in \mathbb{R}^{s\times t}$ 为事件注意力矩阵. 对于任一事件 X_{ijk}, 其不同模态子事件共享注意力 A_{jk}. 考虑到人类处理数据的能力是有限的, 针对单独个体, 假设其在某一时空范围的注意力总和为 1. 当接收到的事件信息已知时, 个体在感知环境的过程中会不断调整有限的注意力, 以达到信息量最大化:

$$K^* = \max_{||A||_1=1} \left\| A \odot \sum_{i=1}^{m} I_i \right\|. \tag{4}$$

从式 (4) 中可看出, 当注意力集中在模态密集的时空事件时, 获取的信息量达到最大值. 因此, 个体可以利用多模态时空数据获取更多信息.

近年来, 注意力机制在计算机视觉、自然语言处理等领域的广泛应用, 证明了对特定事件的关注有助于提高机器的学习能力, 而多模态学习的成功也印证了多模态时空数据联合方面的优势. 因此, 式 (4) 中的模型可以尝试解释多模态认知计算的内在机理, 刻画机器从数据中提取信息的过程. 然而, 是否获取到的信息量越大, 机器就越接近人类的认知水平? 当前, M6 [6], CLIP [7] 等通用多模态学习模型已经在特定任务上取得了接近于人类的效果. 如表 1 [6~13] 所示, 这些模型往往需要千万级的训练数据, 与人类认知能力还存在很大差距. 将单位数据的信息提供能力定义为 “信容” (information capacity) [14], 与此对应, 机器的认知能力即为从单位数据获取最大信息量的能力:

$$\rho = \frac{K^*}{D} = \frac{\max_{||A||_1=1} \|A \odot \sum_{i=1}^{m} I_i\|}{D}, \tag{5}$$

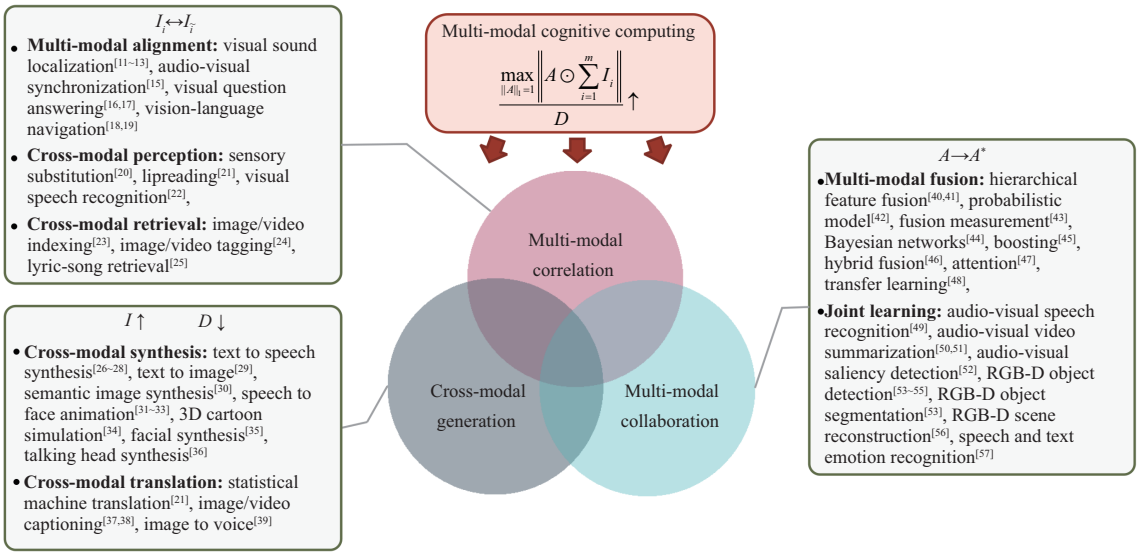

图 1 (网络版彩图) 多模态认知计算任务关系图

Figure 1 (Color online) The relationship among the fundamental tasks in multi-modal cognitive computing

其中 D 为事件空间 X 的数据量. 因此, 可以从三方面提升机器的认知能力: (1) 优化 A, 即使机器获得更大信息量; (2) 增大 I, 即利用对于给定任务信息量更大的数据; (3) 减小 D, 即减小数据量. 利用尽可能少的数据实现信息量的最大化, 即代表了更强的认知能力. 因此, 本文围绕以上 3 个共性关键问题, 以提升机器认知能力为核心, 对多模态关联、跨模态生成和多模态协同 3 个基本任务进行了梳理, 如图 1 所示. 具体如下所述.

(1) 多模态关联 [11~13, 15~25] 是提高 ρ 的基础. 它通过挖掘不同子模态事件在空间、事件、语义层面的内在一致性, 将子模态事件映射到统一的信息空间, 实现多模态的对齐、感知与检索识别. 通过多模态关联, 可以挖掘不同模态间的对应关系, 以进一步提升认知能力.

(2) 跨模态生成 [26~39] 通过增大 I, 减小 D 来提升 ρ. 它将信息以模态为载体进行传输, 利用不同模态的差异性, 对已知信息进行跨模态的合成与转换. 在跨模态合成中, 利用更加直观, 易于理解的模态对信息进行丰富和补充, 增大 I. 在跨模态转换中, 寻找更加简洁的表达形式, 在保留信息的同时, 减小 D, 以此提升信息获取能力.

(3) 多模态协同 [40~57] 通过优化 A 以实现信息量 K 最大化. 它利用不同模态间的关联与互补, 探究高效、合理的模态间联合机制, 优化 A. 通过学习以图像、视频、文本、语音为代表的多模态数据的一致性表达, 实现信息的融合与增强以提升在任务 T 上的性能.

反观人类认知, 认知的提升离不开对现实世界的联想、推理、归纳与演绎, 与多模态认知计算中的关联、生成、协同对应. 本文将人类与机器的认知学习统一为提高信息利用率的过程. 随着人工智能的影响逐渐深化, 多模态认知计算的研究向深度和广度飞速拓展. 作为多模态认知计算的三条主线, 多模态关联、跨模态生成和多模态协同是提升机器认知能力的有效途径, 已成为国内外科研人员密切关注的研究热点. 本文对相关工作展开详尽的调研和介绍, 系统性地梳理了多模态关联、跨模态生成和多模态协同的历史沿革和发展现状, 深入地讨论了多模态认知计算领域面临的机遇和挑战, 并对其未来的发展方向和路径进行了思考与展望.

本文的组织框架如下: 第 2 节, 介绍了多模态关联任务的发展现状, 分为多模态对齐、跨模态感

知和跨模态检索 3 个部分, 并进行分析与讨论; 第 3 节, 介绍了跨模态生成任务中的跨模态合成和跨模态转换方法, 并进行分析与讨论; 第 4 节, 从模态融合和联合学习两个方面介绍多模态协同任务, 并进行分析与讨论; 第 5 节, 对多模态学习面临的挑战和未来发展趋势进行探讨与展望; 第 6 节, 围绕多模态认知计算中的开放问题展开设想; 第 7 节, 对全文进行总结.

2 多模态关联

多模态感知与学习, 通常是通过对同一个实体或时空事件在不同模态空间内予以阐述或描述, 从而得到不同模态的数据. 例如, 采用 RGB-D 相机对同一场景进行拍摄而得到 RGB 彩色图像描述和 Depth 深度距离描述; 采用摄像机对说话人进行语音采集得到其说话内容的语音信息和相对应的唇部运动信息, 这些多模态描述能够更全面地刻画同一客观实体的多维度信息, 从而提升模型的理解与认知能力. 不同模态在表征同一客观实体时所能获得的信息量是不同的, 例如, 在上述对说话信息表征时, 语音获取的说话内容信息量一般要高于从视觉唇部获取的信息量. 虽然不同模态所获得的信息量是不同的, 但是因为它们表述的是同一客观实体, 因此其所获得的信息是存在较强关联关系的, 如发出不同的音素时, 其唇部的视觉运动表现是不同的. 因此, 为了有效刻画多种模态信息间的关联, 需要对不同模态所获得的信息量进行有效分析与对齐, 进而实现高质量的多模态感知与学习. 即在对不同模态所获取的信息量进行联合感知求和基础上, 需进行高质量的信息关联与对齐, 从而为后续的多模态感知与检索奠定基础. 例如, 对于模态 i 和 $\tilde{i}$, 基于不同模态所获得的信息量, 通过特定函数 $f(\cdot)$, 实现不同模态信息量的关联对应, 即

$$f(h(X_{ijk}, T)) = f(h(X_{\tilde{i}jk}, T)). \tag{6}$$

通过优化不同模态所获取信息间的关联目标 $f(\cdot)$, 实现不同模态间关联关系的获取. 本节从多模态对齐、多模态关联和多模态检索三方面阐述多模态关联相关工作. 其中, 多模态对齐是一类基础性需求, 如图像区域内容和文字词汇的语义对齐, 视觉唇部运动与语音声素之间的时间对齐等. 在对齐的基础上, 通过最大化模态间关联满足多模态感知、检索等实际任务需求.

2.1 多模态对齐

实现多模态感知与学习的前提是能够对这些多模态数据进行有效的对齐. 即在时间与空间上, 发现不同模态描述内容的对应关系并进行相互间的关联与匹配, 从而更好地为后续的多模态认知服务. “对齐” (alignment) 这一概念不仅仅出现在多模态认知计算中, 在单模态图像数据研究中, 还存在着与之相类似的概念, 即配准 (registration). 图像配准是图像处理研究领域中一类典型的技术问题, 它的目标在于将同一对象在不同拍摄条件 (如不同的拍摄视角、光照条件等) 下获取的图像进行比较或融合. 具体而言, 假设给定拍摄同一场景的两张图像, 图像配准希望寻找一种空间变换准则将一幅图像的内容对齐到另一幅图像上, 使得两幅图中相关的空间位置的特征点一一对应, 从而实现上述拼接或检测等任务目标. 因此, 图像配准本质上是单模态图像数据内部, 从一个图像到另一个图像在内容特征或像素上的对齐. 相比之下, 多模态对齐任务需要实现的是不同模态在内容或模态特征上的对齐. 例如, 给定一幅图像和一段描述文字, 要找到图像中哪些区域与句子中的哪些词汇相对应. 因此, 不同于单模态数据对齐, 多模态对齐需要首先缩减模态表征上的鸿沟, 通过对不同模态内容的有效建模与获取, 从而实现在关键表征层级 (如语义层级) 上的对齐.

通过分析与对比现有多模态对齐相关的工作, 本文将其划分为 3 个类别, 如图 2 所示, 分别从空

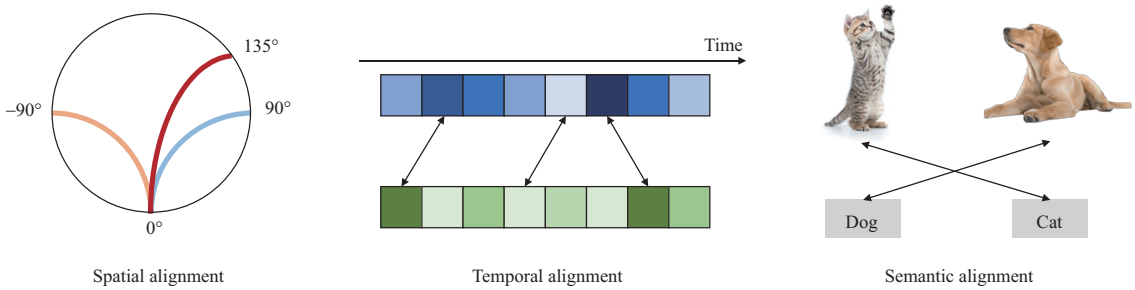

图 2 (网络版彩图) 多模态对齐示意图

Figure 2 (Color online) Illustration of multi-modal alignment

间对齐、时间对齐和语义对齐等 3 个角度出发, 挖掘不同模态内容在物理时空与高层语义层级的关联关系, 从而更好地为后续多模态认知提供有效支撑.

2.1.1 空间对齐

多模态空间对齐一般是指以不同模态在空间位置上的一致性或关联性为基础进行模态内容对齐的多模态对齐方法. 在大多数情况下, 不同的模态数据在空间维度上拥有与模态属性相关的分布特征, 如视觉图像表征视觉物体在二维空间中的分布, 深度图像记录物体在三维空间中的位置信息, 声音在不同空间场合内拥有不同的混响特性等. 虽然在空间中不同模态具有差异化的表征与分布特性, 但是考虑到它们是对同一空间在不同模态形式上的刻画, 其隐式地蕴含不同模态在空间上的一致性关联. 这种空间一致性关联提供了多模态空间对齐的必要基础, 并可为不同模态在空间关联上的构建与学习提供保障.

例如, 基于多模态空间对齐的一类代表性的多模态任务是空间声源定位 [58]. 该任务在给定某一声音模态信息后, 需在对应的视觉场景中定位对应的发声物体, 即实现不同模态在空间上的对齐. 为实现有效的空间对齐, 提升声源空间定位质量, 当前的工作已从传统单通道声音与二维平面图像中物体的一致性关联 [59], 逐步拓展到可提供丰富空间信息的多通道音频和图像内发声物体间的对应关系 [60]. 得益于空间音频信息提供的丰富信息, 视音模态间的空间对齐质量得到了显著的提升. 最近的一些研究通过搭建丰富的视音频空间信息采集设备 [61], 进一步将视觉深度信息引入到空间声源定位任务中, 实现了具有深度信息的三维空间视觉和空间音频在声源位置上的对齐, 从而极大地提升了复杂视音场景下的声源定位表现. 在空间声源定位任务中, 除了引入更多的空间信息来提升空间对齐质量外, 研究者在空间对齐关系的推理上提出若干行之有效的方法. 一些工作将同一视频内的视音频信息作为正确的空间匹配关系, 而源自不同视频的视音频则作为错误的空间匹配关系, 从而得到人为构造的空间对齐关系的正负样例让多模态模型进行判别性学习 [62], 也有工作将音频同局部视觉空间的相似度作为空间对齐的标准考量, 即引入空间注意力机制对多模态空间对齐进行推理学习 [63]. 在最近的一些研究中, 考虑到多模态空间可能存在多个发声物体, 研究者提出在全景空间中通过对视音信息进行空间划分的方式构建正负样例, 实现更精细化的空间对齐 [15].

受空间对齐在空间声源定位任务上的启发, 近些年研究者提出了多样化的基于多模态空间对齐的学习任务, 如空间音频驱动下的深度图生成 [64], 空间一致性约束下的多模态自监督学习 [65], 以及视觉信息引导下的空间音频生成 [66]. 这些多模态学习任务以空间对齐为基础, 通过利用模态间的空间对应关系, 实现了多模态空间内容上对齐, 进而在真实数据上展示出了非凡的表现. 空间对齐作为一类基础性的多模态感知能力, 对智能体在物体与环境理解表现上起着至关重要的作用. 当前的空间对齐

研究大多集中在某些特定任务场景之中, 并未将其抽象为一类可以使用到泛化问题的能力, 但这却是探究多模态认知机理的核心问题之一, 因此在未来一段时间中需要给予更大的关注.

2.1.2 时间对齐

与多模态空间对齐不同, 多模态时间对齐一般是指以不同模态在时间尺度上的对应关系为基础进行模态内容对齐的多模态对齐方法. 对于不同模态信息在时间尺度上的对齐关系, 认知神经科学家在哺乳动物身上发现了有趣的现象. 在麻醉状态下, 某些位于猴子大脑颞上沟部位的细胞会同时对不同感觉 (如视觉、听觉等) 的刺激起反应 [2], 尤其是当不同模态的信息在时间上同步时, 这种联合感觉的刺激就会被增强. 而类似的现象也会发生在诸如大脑中上丘等位置. 随后, 研究人员将注意力转移到人脑, 尝试研究人脑在处理不同模态信息时是否也对时间尺度上的对齐关系敏感. 在 1997 年, 研究者使用功能性核磁共振设备对唇语理解这样的行为进行了分析, 他们发现在大脑左侧半球颞上沟的一个区域会在唇部运动和声音信息匹配时表现得更活跃, 而在不匹配时则会更不活跃. 受上述认知神经科学实例启发, 不同模态信息在联合感知时, 其在时间尺度上的对齐关系对于理解多模态内容而言起着至关重要的作用. 进一步地, 对于机器学习研究者而言, 通常期望在进行多模态感知时, 提升模态间的时间对齐关系, 从而提升多模态关联与感知的质量.

将不同序列在时间尺度上进行对齐是一个经典的机器学习问题. 在多模态关联场景下, 需要首先对不同的模态内容有精准的建模与表征, 进而结合现有的序列对齐方法实现多模态时间对齐. 在早期研究中, 多模态时间对齐主要聚焦在视音频语音识别领域, 通过借助典型关联分析 (canonical correlation analysis, CCA) 方法, 学习视音模态在不同时刻的最大化相似度来实现多模态时间对齐 [67]. 与此同时, 部分研究者也提出采用动态时间规整 (dynamic time warping, DTW) 方法为不同模态的对齐提供求解策略 [68]. 不同模态序列在特征提取后, DTW 目标是求解不同序列匹配间对应的平方和损失的最小解. 考虑到上述问题指数级增长的计算复杂度, 研究者提出采用动态规划对其进行求解 [69]. 在上述将多模态内容建模与序列匹配方法相结合的两类方法基础上, 多模态学习领域的研究者也在尝试从多模态数据本身固有的性质出发, 提出基于对齐学习的多模态关联学习观点. 在这种观点下, 研究者假设存在的多模态数据是部分已对齐的或者在较为粗粒度层级是对齐的, 通过引入例如三元损失 [16], 对比损失 [18], 或基于排序的损失 [70] 等度量学习相关的目标实现不同模态间的对齐关系学习, 以实现对不同模态序列间的时间对齐预测 [16] 或精粒度对齐推理 [71].

根据上述内容, 时间对齐是启发生物体的一类多通道知觉现象, 但是当前大多数研究侧重如何实现高质量的时间对齐, 而对时间对齐产生的原因缺乏相关的研究. 本文认为, 从机器的认知能力上来看, 即式 (5), 时间对齐能够为在减小数据量 D 的基础上, 实现提取信息量的最大化, 而信息量的最大化可以显著提升多模态的认知能力. 同时, 对于需要时间对齐的原因分析会进一步帮助研究者设计更有效的时间对齐算法.

2.1.3 语义对齐

多模态语义对齐一般是指以不同模态在抽象语义空间上的对应关系为基础进行模态内容对齐的多模态对齐方法. 相较于多模态空间对齐与时间对齐, 语义对齐一般面向诸如自然语言等非自然模态[2)], 目标是构建自然语言与其他自然模态之间的对应关系. 例如, 当用一段话来描述一张图片的内容时, 人类可以知道二者在整体内容描述上是相似的, 进一步地, 在诸如物体层级、位置关系层级上对齐单个词语与具体的图像局部内容, 能够实现更为细粒度的语义对齐关系. 当建立自然语言同自然模态

2) 此处自然模态指图像、视频、文本、语音等未经过如自然语言等人为创造的模态.

信息间的语义对应关系后, 这会显著提升对自然模态信息的内容理解, 以及对未来相关任务的执行与预测.

近几年, 虽然多模态相关的学习任务层出不穷, 展现了百花齐放的姿态, 而其中的一大类多模态任务是探究各类自然模态信息同人类自然语言之间的交互与学习, 例如, 看图说话 [72], 针对图像或视频内容的问答 [73], 基于自然语言指令的视觉导航 [19] 等. 通过分析与对比这些同自然语言相结合的多模态任务, 可以发现影响多模态任务表现的部分因素是自然语言同模态信息在语义内容上的对齐. 为此, 近些年众多的相关研究者着力探究实现高质量的多模态语义对齐方法. Owens 等 [74] 指出视觉目标的发声动作和音频信息需要具备同步性才能给人类真实的感官体验, 由此提出了 "视听同步" 任务. 该工作中训练一个多感官特征融合的表征模型, 利用融合特征来反向对齐视频帧和音频信号. 这种对齐技术可以进一步应用于发声视频的声源定位 [65]、发声动作识别 [75], 或多声源分离 [20] 等任务中. Karpathy 等 [10] 在看图说话任务中认为, 在一对给定的图片和对其的文本描述中, 其本身建立很强的关联前提是文本中的词汇和图片的局部内容是对应或关联的. 换而言之, 他们认为这些局部对应的词汇 – 图片内容应当具有较高的相似度, 这种局部相似度可作为整张图片和文本描述间相关性的参考. 因此, 基于上述观点, 相关研究者提出通过学习全局整体性的图片 – 文本关系可以有效推理局部的图片内容和词汇对应关系, 从而实现细粒度的跨图像 – 文本语义对齐 [76]. Li 等 [77] 对图像像素级语义理解进行了详尽的介绍. 进一步地, 研究者也提出将局部图像内容与词汇间的隐式对应关系作为多模态关联中局部需要注意的区域, 并可以此提升看图进行说话描述的文本生成质量 [21]. 与看图说话任务相类似, 在针对特定图像或视频内容的文本问答任务中, 研究者提出通过学习局部文本问题与图像的特定区域之间的相似度, 可以有效推理出问题答案所在的图像局部区域, 从而预测得到正确的答案 [19]. 因此, 通过不同模态局部内容间的相似度可以帮助推理其语义关系, 从而实现细粒度的语义对齐, 并实现相关任务上的性能提升. 在近两年, 同文本相结合的多模态领域也在提出一些新的学习任务, 如视觉 – 语言导航 [19], 视觉 – 语言大规模自监督预训练等 [7]. 相较于看图说话与问答任务, 这些任务 (尤其是视觉 – 语言导航) 同样需要有效的细粒度文本 – 模态内容的对齐关系, 这对于任务执行质量具有可预期的保证. 为了实现这一点, 研究者提出在视觉 – 语言导航任务中, 可以在其他相关任务 (如看图说话等) 上进行预训练, 学习文本与场景内容间的局部语义关联关系, 进而提升基于环境内容的语言指令的理解质量以提升导航效果 [78]. 因此, 对于多模态关联任务而言, 实现有效的模态间语义对齐 (尤其是细粒度语义对齐) 是其中较为关键的组成部分. 但是不同模态的自有属性不同, 其语义组织方式也存在一定程度差异. 当前的研究主要聚焦在高层级的语义对齐方式上, 却往往忽略了不同模态自身的差异化语义组织方式. 如何在差异化的模态语义知识基础上进行模态间的高效对齐与组织等问题可能是未来的潜在研究课题之一.

2.2 跨模态感知

人类通过不同的感官可以获取外界环境的多样化信息, 如视觉感受色彩与运动、听觉感受节奏的律动、味觉感受食物的酸甜苦辣等. 这些多样化的感官信息帮助人类实现对外界环境更为准确的理解. 但是, 随着认知神经科学家研究的深入, 包括人类通过观察自身对外界的感知, 可以发现, 人类不仅可以通过某一模态感受特定的信息 (如视觉感受色彩与运动), 同样可以跨越感官的差别, 用一种感官 "感受" 另一种感官下的信息, 其中一个具有代表性的实例就是唇读 (lipreading). 人类交谈时可以通过聆听对方的说话内容实现彼此间顺畅的交流. 但很多时候, 也会下意识地观察对方的唇部运动. 尤其是在环境中存有噪声时, 仅通过声音不容易听清对方的说话内容时, 观察对方说话时的唇部运动则会显著提升对说话内容的准确理解. 在更为极端的情况下, 仅通过视觉观察说话时的唇部运动而不

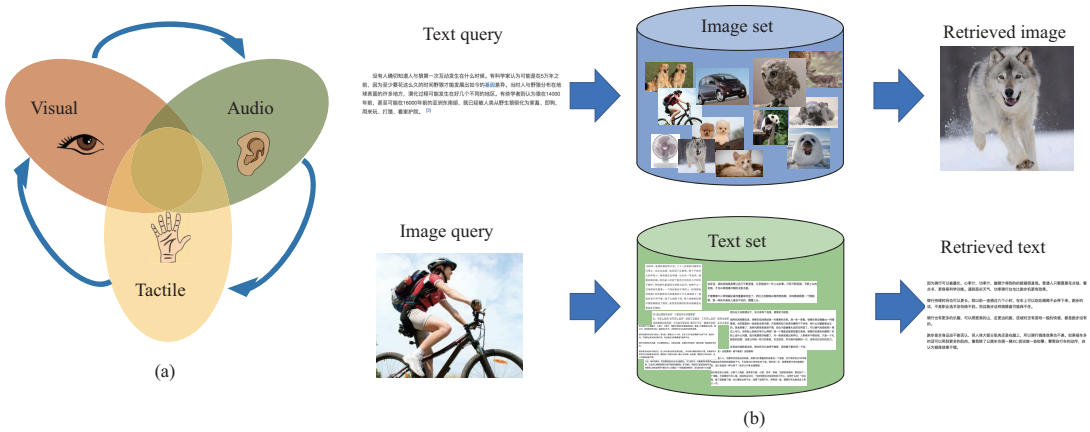

图 3　(网络版彩图) (a) 跨模态感知与 (b) 跨模态检索示意图

Figure 3　(Color online) Illustration of multi-modal (a) perception and (b) retrieval

聆听语音, 即通过视觉理解说话内容, 对于经过特定训练的人群是完全可以实现的, 这种能力被称为唇读. 唇读现象是一类典型的跨感官感知能力体现, 而支撑实现这种能力的基础是说话时唇部的运动与语音内容存在高度的相关性, 如发出 “啊” 的声音时, 嘴部是张开较大的姿态, 而发出 “是” 的发音时, 嘴部是持较小开口且上下唇向外翻折的状态. 从另一个角度讲, 当不同模态间的内容存在不匹配时, 也会对联合感知或跨模态感知的表现产生影响. 例如, 在 1976 年, 研究者在为一段视频配音时发现, 当为视频画面中的唇部运动配不同的说话音素 (phoneme) 时, 研究者会听到额外的第三种音素, 它既不是视觉画面中的那个也不是配音的那个音素 [79]. 这种现象被称作 McGurk 现象, 它是指在不同模态内容不匹配时会产生不同于既有内容的新信息, 从而对客观内容的理解产生偏差. 因此, 通过上述实例可以发现, 有效建立不同模态内容间的关联与对应是实现可靠跨模态感知的重要基础之一. 在更为一般的任务场景中, 跨模态感知也广泛存在于感官缺失的残障群体中. 例如, 视障群体的残疾人士无法通过眼睛获取外界的画面信息, 借助大脑可塑性机制 [22], 可以佩戴额外的感知替代设备 [80] 利用其他感官形式对视觉信息进行感知, 如通过聆听设备编码的声音实现对外界视觉信息的感知 [81]. 借助功能性核磁共振等设备, 认知科学研究者发现, 在没有外界视觉刺激输入, 并在跨模态感知机制的驱动下, 脑部处理视觉信息的皮质区域在执行相关任务时是处于激活状态的 [82]. 因此, 跨模态感知不仅可以在感官正常的群体中提升执行同外界交互等任务的表现, 同时也能够极大帮助感知障碍的残障群体提升环境适应能力.

在认知学家发现人类具备跨感官感知机制及其所带来的感知增强收益后, 人工智能相关研究者思考可计算模型是否也能够具备类似的跨模态感知能力, 如图 3(a) 所示. 考虑到唇读是一类直观且具有代表性的跨模态感知的实际体现, 研究者首先聚焦通过可计算模型实现唇读任务, 即输入模型唇部的视觉运动画面, 然后输出预测的对应说话内容 [83]. 根据前述讨论, 实现有效的跨模态感知需要可靠的模态内容匹配为前提. 为此, 在早期的唇读研究中, 研究者聚焦简单的英文字母唇读识别, 在对单个字母发音时, 用摄像机捕捉唇部的运动细节. 为了准确建模唇部运动和对应发音字母之间的关联, 研究者提出对唇部的轮廓形状进行建模 [84], 或同时辅以对颜色表观的建模 [39], 从而获得特定字母的唇部形状统计信息以实现高质量的多模态匹配. 在此基础上, 研究者也不断将唇读场景从简单的字母识别, 拓展到词语与句子识别 [23]. 近年来, 伴随着深度学习技术在表征学习上展示出了非凡的性能, 研究者提出将唇部画面输入到网络模型中, 并对说话内容直接进行学习与预测 [24]. 为了在唇读识别中有效匹

配其运动信息同音素之间的关系, 研究者提出联结时序分类损失 (connectionist temporal classification, CTC) 并同神经网络相结合以实现端到端训练, 并实现了对人类唇读表现的超越 [85]. 为此, 基于数十年对机器唇读问题的持续深入研究, 研究者通过可计算模型在具体任务上实现了跨模态感知, 为多模态关联学习奠定了新的里程碑. 在此基础上, 研究者也逐渐提出更为多样化的跨模态感知任务, 如基于视频内容的声音生成 [86], 声音驱动下的运动生成 [87], 以及空间音频引导下的深度感知 [64] 等. 在这些任务中, 研究者假设某一模态的信息是部分或完全缺失的, 希望通过设计有效的跨模态感知手段实现从已有模态信息到缺失模态信息的重构与感知, 从而更好地完成特定任务. 以基于视频内容的声音生成为例, Owens 等 [86] 在 2016 年提出对一个给定的静音视频预测生成其对应的声音. 在从视觉到声音的跨模态感知与预测任务中, 他们首先利用卷积神经网络模型刻画视觉动作的动态特征, 随后将获取的视觉特征输入音频序列网络的生成模型以生成对应的声音, 从而完成基于视觉内容的音频生成任务. 因此, 实现跨模态感知的关键在于对已知模态的有效建模, 对未知模态的合理预测, 以及二者在隐语义空间的高质量匹配. 以上任务的成功验证了人类认知对多模态认知计算的指导作用.

在实现诸如唇读、缺失模态生成等跨模态感知的任务后, 研究者进一步提出利用机器辅助的手段, 协助提升残障群体在实现跨感官环境感知时的能力. Hu 等 [88] 提出用于视障等群体的感官替代设备存在算法更新代价高昂的问题, 即需要雇佣大量的视障群体对设备中算法的有效性进行测试评估, 他们提出用可计算模型对视障群体的跨感官感知进行模仿, 从而实现对设备中的算法质量的评价, 以减轻上述设备迭代更新成本. 在未来, 跨模态感知的主要应用场景将不再局限于残障人士的感知替代应用上, 而是将更多地同人类的跨感官感知相结合, 提升人类的多感官感知水平.

2.3 跨模态检索

面对源自不同种类的多样化模态输入, 如何联系并查找其间相似的内容或表征是多模态关联的主要任务, 而以此为代表的多模态任务即为跨模态检索, 如图 3(b) 所示. 跨模态检索是在多模态关联这一研究主题上, 期望跨越不同模态的表征差异, 在语义等表征层级实现模态间内容的有效关联与查找. 由于跨模态检索是现实场景中具有较大需求的一类应用驱动型研究 (如, 在商品检索中以文本检索商品图片内容, 在媒体内容浏览中以词汇短语检索视频片段等), 故其所面对的源自真实多模态场景的数据具有较大多样性和复杂性, 并在检索算法的实现效率上有更为严苛的要求.

面对上述跨模态检索在差异化模态内容关联上的显著特性, 研究人员认为, 不同于单模态检索, 跨模态检索需同时实现单模态内语义相似性的学习以及模态间语义相似性的对齐, 从而实现有效的多模态语义关联与检索. 基于上述的考量, 研究者从不同学习角度对跨模态检索进行了较为深入的研究, 具体可以划分为以关键词为核心的检索方法 [89] 和以模态关联学习为基础的检索方法 [90]. 在早期的跨模态检索研究中, 为了构建有效的模态间关联, 获取自场景内的自然模态信息, 首先用特定的词汇对其中的关键内容进行描述, 以形成对自然模态信息的抽象语义表征, 进而同检索输入词汇或短语在同一语义空间进行关联或匹配以实现检索. 这种基于关键词的跨模态检索方法本质是将模态信息内容建模与模态间关联划分为两个阶段, 因此在检索效率与关键词语义匹配上具有较为明显的优势, 但是分离的两个学习阶段并未有效实现面向模态高质量关联的内容建模与学习, 从而在检索质量方面仍有较大的提升空间. 在近些年的研究中, 研究者提出了面向多模态语义关联的模态语义内容建模, 即将不同模态信息的建模与模态间的对齐关联在同一个学习目标下实现. 从此模态关联学习观点出发, 研究者提出了以度量学习为核心的方法, 即期望内容相关的不同模态信息在模型学习后其表征仍更为相近, 而内容差异过大的不同模态表征在表示空间中相对较远 [91]. 在具体的相似度量方案中, 研究者分别提出以不同模态对之间的语义差异化为目标 [92], 以检索排序质量为目标 [93], 或直接以学习语义相

似性标签为目标的学习方法 [94]. 上述这种以相似性检索为目标的多模态内容学习能够显著提升模态内容的学习质量, 提升跨模态检索效果.

跨模态检索是一类面向真实场景需求的应用研究, 由于需要面对日益增长的数字化内容, 高效的检索算法成为近些年跨模态检索的一大研究重点. 为了实现在提高检索效率的同时不降低检索质量, 研究者提出通过压缩模型或模型小型化以降低模态内容的建模计算代价 [95], 或通过将模态内容的语义表征进行二值化以提升关联匹配的效率 [96]. 在近些年, 因为上述后者在检索效率与质量上的优势, 故其渐渐作为提升跨模态检索效率的主流方法. 一般地, 研究者将二值化的模态表征编码称为哈希 (hashing) 编码, 并把将实质模态表征到哈希编码的过程称为哈希投影. 因此, 实现有效的模态表征二值化的关键在于哈希投影. 与上述方法类似, 研究者以度量学习为核心, 以满足模态内语义相似性和模态间语义一致性为目标对二值编码进行学习 [97]. 最终, 通过对不同模态内容的二值化编码进行汉明距离 (Hamming distance) 度量, 即可实现高效的跨模态检索.

近些年, 随着数字媒体采集设备的普及, 数字化的媒体内容正急速增长, 并由此带来了多样化的跨模态检索需求, 如脚本指导下的动作事件定位 [98] 与视频摘要 [99], 图像 – 歌曲检索与推荐 [12] 等, 这些多样化的跨模态检索任务对模态内容的精细化理解和高质量关联提出了新的挑战, 也将是未来多媒体内容分析的研究重点.

2.4 分析与讨论

多模态关联本质上是在探究并挖取源自不同模态的内容在不同层级的关联对应关系, 并将这种关联对应关系用于诸如模态间交互等相关的任务之中. 因为模态信息的多样化以及不同模态组合形式的不同, 多模态间的关联关系可能处于不同的层级或角度, 如上述介绍的时间层级、空间层级与语义层级. 在实际中, 模态间关联也有可能同时处于上述的不同层级, 这需要面向具体任务需求进行针对性探讨, 如面向空间感知的声源定位需实现模态间的空间对齐, 而视频内容理解等任务通常需要实现不同模态在时间尺度上的对齐. 受认知科学中对大脑跨模态感知研究的启发, 研究者发现通过学习并构建不同模态信息间的关联关系, 可以显著提升多模态相关任务上的表现.

因此, 在学习并构建不同模态信息间有效关联的基础上, 研究者提出通过跨越模态的表征差异, 实现某一模态对另一模态内容的感知. 达到上述跨模态感知的根本基础在于模态间信息表征的高质量对齐, 例如, 唇读任务的表现依赖于声音音素与唇部运动的对齐. 基于对齐的模态间信息表征, 可以实现从已知模态的表征对未知或不可见模态内容的生成与感知. 进一步地, 在跨模态感知的基础上, 多模态关联可以更为有效地面向真实场景的应用需求, 如多模态信息检索. 由于跨模态检索需实现单模态内语义的学习和不同模态间语义相似性的对齐, 因此面对真实环境下复杂且多样化的模态, 跨模态检索是对模态间关联关系学习质量的一次验证. 随着当前数字采集设备价格的日渐亲民化和互联网的高速发展, 数字模态内容的大量增长对多模态关联学习也提出了新的挑战. 面对更大的数据规模和多样化的模态内容关联形式, 需要提出更为高效的关联学习方法.

3 跨模态生成

正常状态下, 人类的多通道感知和中枢思维系统使其具有天然的跨通道推理和生成能力. 例如, 阅读一段小说情节时脑海中会自然浮现相应的画面. 参考这一现象, 本文将多模态认知计算中跨模态生成任务的目标定义为赋予机器生成未知模态实体的能力. 传统的机器生成任务通常在单一且固定的模态上进行, 例如, 利用已知的一段对话生成后续情节 [25], 或利用已有图像合成一张新的图像 [100].

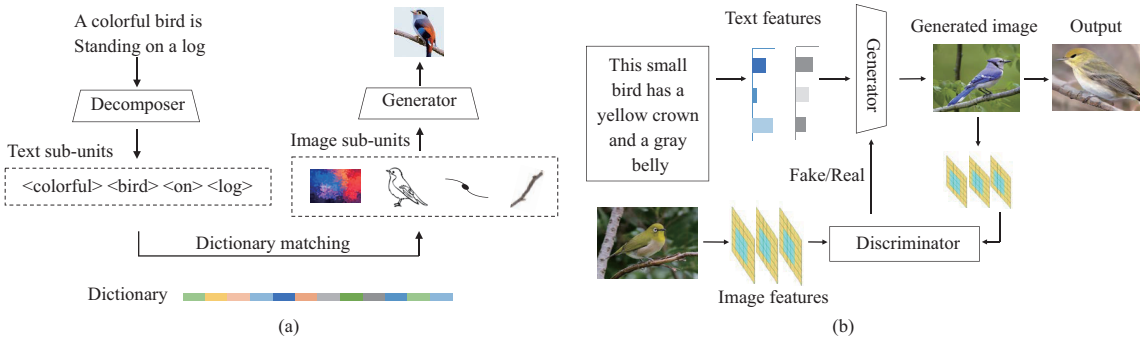

图 4　(网络版彩图) 典型的两种跨模态合成方式

Figure 4　(Color online) Two representative strategies of cross-modal synthesis. (a) Instance-based modality synthesis; (b) generative model-based modality synthesis

类比于此, 跨模态生成是涉及多种不同模态信息的实体生成过程, 利用多模态信息之间的一致性和补充性来生成新模态下的事物. 从信息论的角度看, 跨模态生成任务促使不同模态之间通过信息流动, 提升个体在既定时空内可感知的信息量. 假设已知某实体的 m 个模态信息 $\{X_1, X_2, \ldots, X_m\}$, 跨模态生成任务可以概括为

$$X_p = \arg \max_{\|A\|_1=1, X_p} \frac{\|A \odot (I_p + \sum_{i=1}^{m} I_i)\|}{D}, \tag{7}$$

其中 X_p 是缺失待恢复的模态实体. 随着自然语言处理、智能语音、计算机视觉等技术的快速发展, 建立在文本、语音、图像、视频上的跨模态生成任务层出不穷, 例如, 一句话生成图像 [101], 一段场景产生音频 [102] 等. 这些不同模态对信息表达方式的不同, 对信息的传达能力有很大差异. 在绝大多数人的认知世界中, 一定时空和目标条件下文本、音频、图像、视频这些模态信号能传递的信息量是逐渐上升的. 简单来说, 同一事物的声音比文字能直观传达的信息要更加丰富, 而图像相比声音更加直观一些. 大多数情况下, 日常生活中广播比文字要更容易被多数人接纳; 当看到一张狗的照片时就比听到这只狗的声音了解得更多; 而动态的视频带给人类的感受要更加深刻. 但同时, 同一事物用信息量丰富的模态来描述时也占据更大的存储空间, 带来信息处理效率上的负担. 综合两方面因素, 本文从认知计算的角度将跨模态生成任务的本质归纳为在多模态信息通道内提高机器认知能力的问题. 进一步地, 可以将该任务划分为提高信息量 I 和减小数据量 D 两种方式, 即跨模态合成和跨模态转换两大类. 下面详细介绍这两类跨模态生成技术.

3.1　跨模态合成

单一模态的合成问题可以理解为信息的叠加和重构, 而跨模态合成任务则是依据旧模态信息对新模态实体进行具象化的过程, 合成的原则是提高和丰富已有模态提供的信息. 合成的方式是在原有模态信息上加入更多推理得到的内容, 促使从简单模态到复杂模态的实体生成. 近年来, 跨模态合成的研究以语音合成和图像合成技术为代表, 例如, 将文本合成为带有情感的语音或图像 [103], 如图 4 所示.

早期的跨模态合成问题研究主要采用实体关联的方式, 将多模态的实体关系直接存储在字典库中, 根据库信息来匹配得到新模态样本. 这类方法的核心在于如何使用实体关系, 最简单直接的做法是设计跨模态检索算法. 以语音合成任务 (text to speech, TTS) 为例, 将待合成的语音看作基础声音的选择和拼接问题. 先通过检索得到字典库中匹配的文字条目, 在文字 – 声音字典中找到待合成的词

元素, 再通过单元拼接来构造指定的声音内容. 在图像合成任务 (image synthesis) 中, 这种方式也是简单有效的. 例如, 一句话生成图像即找到在字典中与给定描述最相符的文字内容, 再映射到图像库中即可. 这种方式尽管能快速合成指定内容和模态的实体, 但其对检索库的依赖程度往往很高, 同时对于自然度和流畅度这些真实语音特征的实现能力也有限. 考虑到语音信号的时间连续性, 基于统计学模型参数估计方法的连续式生成方法在语音合成任务中取得了大量的关注. Zen 等 [104] 全面对比了不同统计学模型在语音合成中的表现, 指出基于隐马尔科夫模型 (hidden Markov model, HMM) 的语音合成方法可以更好地关注语音逻辑上的连贯性, 生成自然度和连续度更强的音频信号. Anderson 等 [105] 进一步提出基于 HMM 的可视化语音合成的问题, 在合成语音片段的同时生成相应的说话人头像视频, 在表达讲话内容的同时将说话人的面部讲话动作可视化地展示出来. 尽管 HMM 的方法可以对连续的信号进行很好地模拟和预测, 但往往存在过度平滑的问题, 对于特征变化较大的场景处理能力非常有限. 例如, 在音乐合成问题中, 由于音乐中包含的情感信息更加丰富, 同时一段音乐中涉及的节奏和音调变化也更加复杂, HMM 方法很难直接处理其中起伏变化的高低音情节. 利用神经网络作为声学模型训练乐谱特征与声学特征之间的映射函数是另一种统计学参数估计方法, 其相比基于 HMM 的训练模型表现出更强大的特征学习和拟合能力. Hono 等 [26] 的最新研究利用深度网络设计了歌声合成系统 (singing voice system, SVS), 在基础的网络中加入乐谱分析模块、音高模块、节奏控制模块等, 能够根据输入的乐谱合成指定节拍和音调的音乐. Yu 等 [27] 利用注意力机制将发声词与文本内容进行对齐. 通过对文本的上下文语义进行分析, 形成有风格、情感和韵律的发音系统. Ren 等 [106] 提出了 FastSpeech 模型, 解决语音合成的速度问题, 同时提升模型在重复吐词、漏词等方面的鲁棒性, 以及在讲话速率、韵律等方面的可控性.

相对于文字和语音信息单元来说, 人类通过视觉感知到的图像和视频信息更为丰富也更为复杂. 文本与音频信息之间通常在内容上存在简单直接的一一对应关系, 但它们与视觉信息的关联就要涉及不同程度的解译过程, 例如, 对自然语言和图像的理解. 计算机视觉和自然语言处理技术的不断突破也促使二者结合的多模态图像合成任务成为近年来的一大热点. 例如, Wang 等 [107] 尝试从新闻报道中生成带图像的故事情节, 试图依靠文本创造图像. 从文本到图像的生成任务需要注入更多的细节信息, 因此相对于语音合成来说更加困难. 当前的图像生成技术以生成对抗网络为主, 通过判别器和识别器联合训练的方式得到逼真的高质量图像. 基于文本描述生成图像的方法不仅要考虑图像的真实性, 更要考虑图像在内容上的合理性. Mansimov 等 [11] 率先利用深度生成网络设计了根据非结构化自然语音描述信息来合成图像的技术, 对描述信息进行关键目标提取和定位, 并用目标信息来迭代渲染整幅图像. Qiao 等 [103] 设计了从文本到图像合成, 再从图像到文本生成的双向级联网络进行深度特征关联, 这种文本 “重描述” 的监督学习方式对于文本 – 图像语义一致性学习能力突出, 补充了传统生成对抗网络 (generative adversarial network, GAN) 语义挖掘能力不足的问题. 风格信息也是重要的特征, 就像人类的情绪一样, 同样内容在不同情绪表达下就有很大差异. Park 等 [28] 提出了一种联合内容描述和风格描述的图像生成方法, 其中的风格信息可以由任意风格图像来指定. 这种风格属性是一种整体的渲染, 就像给图像加上指定类型的滤镜. Zhu 等 [108] 提出语义级的图像合成任务, 给指定语义的局部图像创造多样化的生成结果. 这种任务在图像合成中加入了高度的语义控制能力, 能够精准分割并生成指定语义的图像内容. Schönfeld 等 [109] 提出在 GAN 的判别器中加入语义分割网络, 在训练阶段利用给定的图像语义分割标签去生成指定场景的内容, 进一步精准提高生成图像的语义相关性. 相比于自然目标图像来说, 人脸图像在神态表情方面有更为复杂的信息, 微小的表情变化往往传递出来的信息差异很大. Di 等 [110] 提出多模态的人脸生成模型, 能够根据面部描述特征 (例如, 性别、表情、配饰等) 同时生成多种模态的人脸图像, 包括可见光图像、轮廓图像、热红外图像、偏振图像等. 利用

语音生成面部视频是很早出现的一类任务, 在影视娱乐等领域涉及很多. 早期的研究 [29] 通过对语音和面部动作进行相关联的高斯过程建模得到两者的转换模型. Oh 等 [30] 在深度学习框架上提出端到端的语音和人脸联合训练模型, 利用百万级数据量的 YouTube 数据训练模型参数, 最终得到对年龄、性别、种族等属性都能自主适应的自然转化算法. 近年, 虚拟人讲话技术逐渐成熟并延伸出 3D 动画合成任务 (speech-to-animation), 给一段指定的内容生成能模拟讲话的 3D 虚拟动画. 迪士尼研究院联合多所著名高校提出了一项研究 [35], 借助深度学习训练从语音标签输入序列到口部动作的映射函数, 形成自动生成在面部和口型上符合指定语音的卡通人物.

3.2 跨模态转换

真实场景在生成新模态实体时除了丰富已有信息的目的之外, 还有一些需要复杂模态到简单模态进行转换来降低数据量的任务. 这类多模态任务在进行不同模态的关联性学习时往往关注重要信息的凝练和表达 [31, 32]. 例如, 给一段冗长的视频进行简要的自然语言描述, 或者给一段视频信息生成与之相关的音频信号等, 如图 5 所示. 下文对跨模态转换任务分类探讨.

视频和图像的自然语言描述是将视觉信息转换成文本描述信息的跨模态转换任务, 可以极大地提升视频或图像信息的在线检索效率. 这类任务是计算机视觉和自然语言处理两大技术结合的代表. 图像描述, 即 “看图说话” 问题 (image to text, I2T) 涉及目标检测与识别、场景理解等复杂的图像解析过程. 早期研究 [33] 将图像描述归纳为三段式任务: 图像内容解析 – 目标语义表示 – 自然语言生成. 这种方式的图像描述对图像中解析到主体、客体、场景等视觉元素和文本语法结构 (例如时间、空间、功能等) 进行关系构图, 同时在语义表示阶段可以与外部知识库连接, 从而充分挖掘视觉元素的逻辑描述关系. Yang 等 [34] 认为直接分析图像元素的方式不太可靠, 提出利用在英文语料库中训练好的语言模型作为先验, 再结合输入的图像元素训练专门的隐马尔科夫模型, 模拟图像描述生成的过程. Elliott 等 [111] 从图像中目标和背景的依赖性分析入手, 设计多层的视觉依赖表示方法来刻画不同区域的关联, 解决复杂目标关系的解析问题. Mao 等 [112] 提出多模态神经网络框架, 用对视觉信息更敏感的深度卷积网络提取图像信息, 同时用对逻辑关系更关注的深度递归网络来预测文本描述词. 通过交互式训练两个子网络, 对图像和文本的关联分析能力有显著提升. 深度学习在视觉分析上的成功致使高质量的自然段落生成成为图像描述任务中更有挑战的环节, 通常使用最大熵模型或递归网络模型作为语言生成器. Devlin 等 [113] 对比了两种语言模型的性能, 并指出指标度量结果与人类自然描述仍存在较大差异. 随着变分自编码器结构 (variational auto-encoder, VAE) 在机器翻译任务上的成功运用 [114], 基于自编码器的图像描述方法逐渐受到关注. 一些研究者提出卷积神经网络 (convolutional neural network, CNN) 与递归神经网络 (recurrent neural network, RNN) 结合的自编码器框架, 对图像进行卷积编码得到内容描述向量, 再通过递归网络解码描述向量来与文本描述进行匹配 [72]. Xu 等 [21] 同时将注意力机制引入到自编码器的解码模块, 抛弃传统单一特征向量衔接的方式, 在生成描述词时在全局视觉信息中逐个寻找注意力区域. 这种注意力机制可以在解码时关注图像的不同局部区域, 从而再提高解码精度. Wu 等 [37] 指出现有自编码器在解码阶段利用特征编码的方式无法充分挖掘图像信息, 继而提出在生成描述时循环且带选择性地调用部分视觉信息, 加强描述生成的准确性. 除自编码器外, 强化学习 (reinforcement learning) [115] 是目前图像描述任务的另一大主流框架. 强化学习是一种通过观察环境来调整行动, 以取得最大化预期收益的学习模式. 图像描述任务在训练时的描述词是同步输入的, 而测试时只能逐个预测, 因而存在损失评估差异. Rennie 等 [116] 在增强学习框架下提出自批判训练方式, 直接利用测试结果来矫正文本预测模型的收益值, 提高文本预测的准确性. Xu 等 [117] 指出单层的策略网络和收益网络无法有效处理复杂多样的文本和图像. 他们提出多层级的强化学习方式, 利

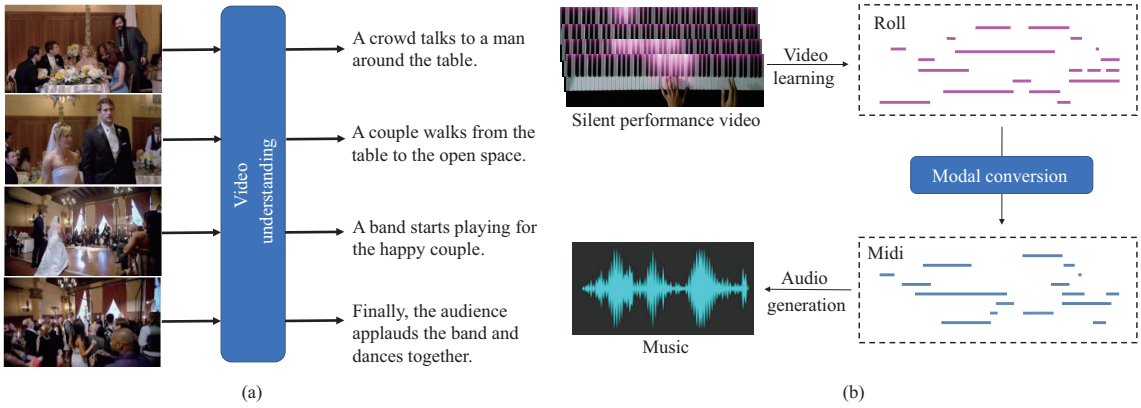

图 5　(网络版彩图) 常见的跨模态转换任务

Figure 5　(Color online) Illustration of cross-modal translation. (a) Video to text description; (b) video to music translation

用联合评估图像和文本收益的策略生成词描述和语句描述. 自然环境下的图像内容相比有限域的图像更为随机和复杂, Guadarrama 等 [118] 解决了面向大规模自然场景数据集的任意活动描述问题. Che 等 [119] 提出视觉关系网络生成更为完备的图像段落描述. 相比于图像描述, 段落描述问题需要对图像进行更加细致和完备的解译, 对目标关系分析的完整性要求更高. Li 等 [38] 提出针对场景图表示来描述目标和目标关系, 对图表示提取视觉和语义表示, 进而输入一个分层的注意力模型来生成单词描述. Wu 等 [120] 提出全局和局部判别模型解决图像描述的结果过于平凡化的问题, 对图像进行细粒度目标分析生成特征更加明晰的描述词. 视频描述是图像描述的进阶任务. 由于视频场景变化更复杂, 目标关系更多样, 逻辑顺序更严谨, 对视频解译和描述的难度也越高.

本文将语音、图像等非文字的表达方式统一称为非自然语言描述. 除对信息进行自然语言转换外, 非自然语言描述类的任务也很常见, 例如, 基于视觉信息的语音生成、视频摘要、语音情绪翻译等. 利用视觉信息合成声音是计算机视觉技术的延伸. 不同的物体在被敲击或撕扯时会发出特定的声音, 声音差异取决于物体的材质以及操作的方式. 基于此, Owens 等 [86] 提出了一种基础的 “看图发声” 任务, 利用 RNN 来预测目标的发声类型, 再结合声音合成模板来生成相应的声音片段. Chen 等 [121] 提出利用深度生成网络来解决视频合成声音的问题. 该工作中对乐器演奏类视频进行分析, 生成符合视频内容的音乐信息. Zhou 等 [122] 提出了自然场景下的视频声音合成问题, 并构建了包含 28109 个片段共计 55 小时自然环境下的视频数据集, 包括室内环境、户外环境等. 衡量算法生成声音的各项性能时通常利用人工方式来评估, 代价十分昂贵. 为此, Hu 等 [88] 提出了一种机器可评价的方法, 将图像生成的声音信号重新转化为图像信号, 再通过图像间的一致性评价来指示合成声音的质量. 电影场景的音轨合成是视频转音频的典型应用, 这类合成声道对于时间同步性的要求也非常高. Ghose 等 [123] 提出了一个全自动的深度学习工具箱为电影场景合成音频轨道, 可以同步相关的音频和视频信号, 或者为关键场景合成或增强音轨信号. 大量的工作将视音任务的关键归纳为高质量的多模态联合表征问题, 利用受限玻尔兹曼 (Boltzmann) 机和深度玻尔兹曼机进行模态联合处理. Hu 等 [124] 指出多模态语义的相似性分析是协助不同模态信息相互理解的关键, 提出基于深度信念网络的多模态语义相似性分析问题, 将语义相似性嵌入特征融合过程. Hu 等 [125] 提出了分层的多模态融合表征方法对多模态信号进行多层融合训练, 相比于传统的单层融合方式能够挖掘更深层的融合信息. 不同于自然语言类的视频描述问题, 视频摘要的任务是在长时间的视频序列上寻找视频关键帧和镜头以压缩和凝练视频信息. Zhao 等 [126] 提出了序列图的重建网络, 使用图卷积神经网络学习视频序列中的长时依赖. Li 和

Zhao [14] 对视频摘要任务的发展和现状进行了详细介绍.

3.3 分析与讨论

生成模型通常包括两部分, 分布预测和样本生成. VAE 和 GAN 是目前主流的两种生成式模型, 在多模态生成任务上也有很多发展. 自编码器 (auto-encoder) 是一种以最小化输入和输出特征差异为目标的无监督特征压缩模型. 这种方式相对于卷积神经网络而言具有更强的解释性, 并非纯粹的拟合. VAE 是一种生成模型, 在编码 – 解码结构上对输入样本和输出样本的分布进行一致性约束, 得到能反映真实样本分布的生成器. 但是, VAE 会依赖很多假设条件, 例如, 隐变量的连续性和低维流形结构. 这种对离散输入进行低维连续性估计的方式, 往往会造成其生成样本比较模糊. 可以从两点来理解, 一是样本中的细节信息通常更多存在于高维特征中, 二是多样化的细节信息很难用单一的低维流形来描述. GAN 模型采用对抗生成的策略, 在生成样本的同时对其性能加以评估, 决定模型的生成走向. 相对于 VAE 来说, GAN 模型可解释性差, 且对抗学习方式很难平衡, 模型优化困难且不容易找到稳定点. 但其可交互的生成策略使生成真实性和清晰度很高. 因此, 对 GAN 和 VAE 的合理结合是生成任务的主要趋势.

多模态生成任务的挑战不仅在于生成质量方面, 更多在于不同模态之间的语义及表示鸿沟, 需要在具有语义鸿沟的前提下进行知识推理. 由于不同模态信息组织的差异, 有效的解析和生成方式也有各自的特点, 例如, 自然语言描述是一种强结构化和规则化的表达方式, 而图像却是一种无结构化数据. 对不同模态特征表达的一致性学习和映射是多模态生成的重要前提. 多模态生成任务的组合方式也各式各样, 对不同任务的关联性分析和跨任务迁移学习也具有重要价值.

4 多模态协同

归纳和演绎是人类认知的重要功能. 人类可以轻松自如地对视、听、嗅、味、触等多模态感知进行归纳融合, 并进行联合演绎, 以做不同的决策和动作. 在多模态认知计算中, 多模态协同是指协调两个或者两个以上的模态数据, 互相配合完成多模态任务. 为了实现更加复杂的任务并提升精度和泛化能力, 多模态信息之间要相互融合, 达到信息互补的目的. 呼应前文, 这本质上是对注意力 A 的优化:

$$A^* = \arg\max_A \left\| A \odot \sum_{i=1}^{m} I_i \right\|. \tag{8}$$

进一步地, 融合后的多模态信息要进行联合学习, 以实现多模态信息对单一模态的超越, 即

$$A^* \odot \sum_{i=1}^{m} I_i \geqslant I_{\tilde{i}}, \quad \tilde{i} = 1, 2, \ldots, m. \tag{9}$$

信息量的增加可以提高单模态任务的性能, 也为开发创新性多模态任务提供了可能.

从生物学角度来看, 多模态协同和人类综合多种知觉作出反应是相似的. 近年来, 随着传感器技术、计算机硬件设备和深度学习技术的更新换代, 多模态数据的获取、计算和应用也变得日新月异. 同时, 建立在视觉、声音和文本等模态上的多模态协同研究也取得长足发展. 本节重点总结了多模态协同中的模态融合与联合学习方法. 其中, 模态融合分为前期、后期和混合融合策略, 如图 6 所示. 联合学习根据其目的不同分为提升单模态任务性能和解决新的挑战性问题. 接下来, 本节将分别进行介绍.

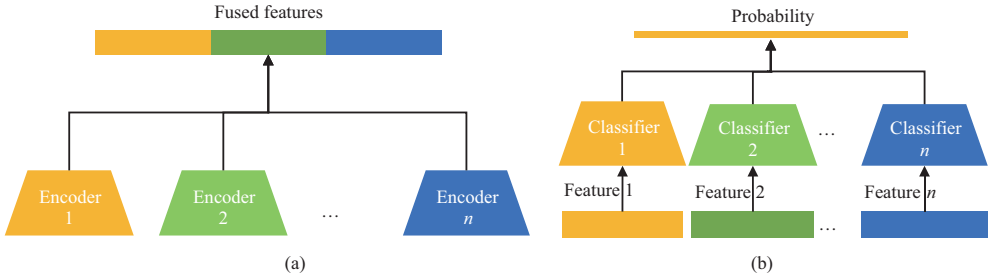

图 6 (网络版彩图) 多模态融合中的 (a) 前期融合与 (b) 后期融合示意图

Figure 6 (Color online) Illustration of (a) early fusion and (b) late fusion in multi-modal fusion

4.1 模态融合

多模态数据之间存在差异, 模态融合过程中需要考虑数据格式、时空对齐、噪声干扰等问题. 当模态数据间的异构性较大时, 模态融合方法就显得尤为重要. 一般来说, 多模态数据在进行融合时, 在语义、时间和空间维度上往往是对齐的. 另外, 不同模态数据的信息既有相关性又有独立性. 例如, 彩色图像和深度图在同时提供物体轮廓信息的同时, 还可以分别反映物体的色彩变化和深度信息, 因此如何对不同模态的特性进行挖掘和建模也是模态融合的核心问题. 此外, 针对不同任务, 每个模态发挥的作用不同, 因此模态融合是任务相关的. 接下来, 本节将以融合阶段和融合方法作为主要内容进行介绍.

4.1.1 融合阶段

多模态数据信息包括视觉、听觉、文本、深度、点云、运动等各类型数据. 多模态数据的融合是对多模态信息进行协同感知、联合学习, 在融合过程增大信息量的同时也实现了信息的互补和熵减效率的提高. 多模态数据按照融合时间和方式的不同可以分为前期融合 (early fusion)、后期融合 (late fusion) 和混合融合 (intermediate fusion) 三类.

前期融合是特征层面的融合, 也称为早期融合, 是融合阶段应用最广泛的策略, 它是将从每种模态的数据中提取到的特征在执行分析任务之前进行融合, 是特征级别的融合. 前期融合的优点是可以在模型输入时即实现多种不同模态数据的信息互补, 从而保证了多模态数据的全流程融合. 但是, 前期融合无法应对不同模态数据的低水平特征中存在的噪声和多模态特征之间的时间同步问题, 这些问题增加了模型的学习难度, 甚至导致一加一小于二的结果. 目前在特征层面的融合方法主要有概率论统计方法 [127]、神经网络方法 [128]、基于特征抽取的方法 [129] 和基于搜索的方法 [130]. 这几类方法没有严格的界限, 各自有优缺点. 在实际应用中, 以上方法常常组合使用. Nefian 等 [131] 建立动态贝叶斯网络用于对视频和音频线索之间的特征相关性和时间相关性进行建模. Li 等 [132] 提出了适用于多模态数据的动态图自编码器, 用于提取图结构数据的特征. Zhang 等 [40] 提出动态图嵌入的无监督特征学习方法. Wang 等 [133] 提出基于 ℓ_1 范数的多视角聚类方法, 以抑制多模态数据中的噪声. Singh 等 [134] 使用 Fisher's Discrimination Ratio [42] 进行特征融合和筛选. Li 等 [135] 提出尺度可调节的无参数二部图融合方法, 对多源特征构成的邻接图进行融合.

后期融合是决策级别的融合. 在后期融合中, 不同模态的数据和特征在不同的模型中进行训练, 并得到各自独立的单模态决策结果. 然后使用决策融合策略组合单模态决策, 得到最终的融合决策向量 [136]. 和早期融合不同的是, 后期融合由于是在语义级别进行决策的, 可以有效规避不同模态特征之间的同步和对齐问题, 且单模态决策通常具有相同的表示方法, 单模态决策的融合相对来说更加容

易和可扩展. 晚期融合的另一个显著优点是, 不同模态的数据是异步处理的, 这就意味着可以在不同的模态上选择最合适的分析和处理方法, 且每一个单模态方法之间是互不干扰的, 这大大提升了系统的鲁棒性和可扩展性 [137]. 同时, 当模型缺少某一个或者某几个模态的输入时, 模型依然可以正常预测. 但是, 后期融合也有一些缺点, 由于使用不同的分类器来获得单模态决策, 各个模态之间的特征往往不能得到有效利用, 且整个学习过程比较固定, 系统也更复杂 [43]. 在后期融合中, 最常用的方法是对各模态的决策进行加权、求和和投票 [137], 也可以使用一些机器学习算法 [41]. Liu 等 [44] 使用 Adaboost [41] 算法进行决策融合.

混合融合, 顾名思义, 是在特征层面和决策级别同时进行的融合方法, 是为了充分整合前期融合和后期融合方法的优点, 规避单一融合存在的一些显著问题, 在混合融合中, 特征层面的融合是和决策级别的融合并行进行的, 在最终决策时同时考虑特征融合的决策结果和决策融合的结果. 混合融合策略是十分有效的, 在许多多模态任务中均取得很好的任务效果. 但其存在的问题也比较明显, 由于同时使用两种融合策略, 模型通常会更加复杂, 训练难度也大大提升. 不过随着深度学习和计算算力的迅速发展, 这一问题正得到有效缓解和解决. Lan 等 [138] 提出了双融合的混合融合方法, 在事件检测任务中有效提取特征的同时, 很好地缓解了过拟合问题. Bendjebbour 等 [139] 在多传感器图像的统计分割任务中, 提出基于 D-S 证据理论 [45] 的混合融合方法, 取得了很好的融合效果. Xu 等 [140] 提出了一个利用多种异步信息和外在信息源的可扩展融合框架, 在足球运动的事件监测任务中展示了混合融合的优异性能.

4.1.2 融合方式

多模态信息融合是多模态机器学习的一个关键的研究方向, 它将从不同的单模态数据源中提取的信息集成到一个紧凑的多模态表示中, 目的是构建一种可以同时处理和互动地关联多种模态信息的模型. 多模态数据的融合方式可以分为基于规则的方法和基于学习的方法. 基于规则的方法包括串行融合、并行融合和加权融合的方法, 这些操作通常只有很少相关参数.

串行融合是将多个模态特征在某一维度进行串联拼接组成一个相同的维度特征, 从而组合为新的高维度的融合特征向量 [141~143]. 此方法的优点是计算简单、容易实现, 缺点则是当融合的特征维度太高会出现 “维度灾难” 问题. 并行融合是利用复向量把多模态特征向量合并, 之后在向量空间对这些特征向量进行融合. 并行融合的方法在计算上虽然繁杂, 但该方法具有更好的鲁棒性. Katsaggelos 等 [46] 阐述了视觉和语音两种不同模态数据的并行融合方法, 同时讨论了语音和视觉数据不同步和只有一种模态数据存在时的研究. 目前的数据融合方法致力于达到建立模态之间高效信息互补的要求. Nojavanasghari 等 [144] 提出在 3 种不同数据模态的高层特征上进行并行数据融合, 然后将融合好的数据输入到神经网络中训练预测网络. 该方法建立在模态数据的高层特征上, 其融合方法使得各模态之间具有信息同步和互补的特点. 加权融合是对不同模态的数据特征取不同的权值进行加权求和 [145]. 但该方法要求预训练模型产生的向量要有确定的维度.

基于学习的方法是指通过学习的方法进行模态数据融合, 包括注意力机制模型、迁移学习和知识蒸馏等. 注意机制 [146] 通常指一组向量的加权和, 这些向量具有标量权重, 是由一组 “注意力” 模型在每个时间步动态生成的. 这组注意力的多个输出可以动态地产生求和时需要用到的权重, 因此最终在拼接时可以保存额外的权重信息. 常见的注意力机制模型又可分为图注意力机制 [147]、图和文本的对称注意力机制 [148] 和双模态 Transformmer 注意力机制 [149]. 为解决多模态数据融合深度模型训练困难和通用性较差的问题, 迁移技术被应用到数据融合的模型中. 迁移技术不仅具有知识迁移能力又具有多模态相互增强的能力, 并可以提高不同模态之间的数据融合能力. Gao 等 [150] 第一次提出将多个

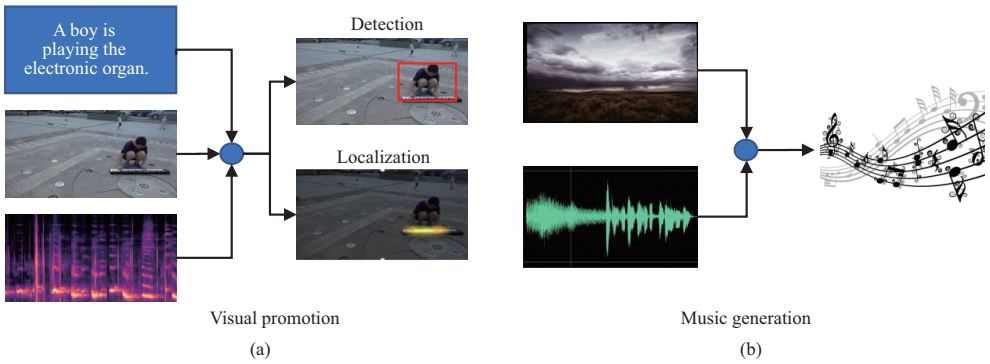

图 7 (网络版彩图) 多模态联合学习中的 (a) 模态性能提升与 (b) 模态创新应用示例

Figure 7 (Color online) Examples of (a) multi-modal performance improvement and (b) new type of multi-modal task

模型结合起来进行迁移学习的模型, 根据模型对每个测试示例的预测能力动态分配权重. 它可以将各种学习算法的优点和来自多个训练域的标记信息集成到一个统一的分类模型中, 然后应用于不同的领域. Moon 等 [151] 提出不同模态之间迁移分类的方法, 文中提出的方法具有的鲁棒性和高适用性, 使得泛化能力更强. 为了解决多模态任务中对计算资源以及计算时间要求高的问题, 蒸馏模型被应用到数据融合的任务中, 蒸馏模型不但可以节省计算资源并且还能保持性能. Jin 等 [152] 验证了知识蒸馏在不同方法的有效性. 此外, Agarwal 等 [153] 还提出了模态对应的知识蒸馏模型, 其每个模态都有不同的预测显著性, 文中为每个模态定义了显著性分数, 学生网络根据显著性分数在每个模态上模仿教师网络的输出.

4.2 联合学习

多模态信息融合完成后需要进行联合学习. 通过模态信息间的联合学习可以帮助模型挖掘模态数据间的关系, 建立起模态与模态间的某种联系. 这种联系可以是辅助性或互补性的. 辅助性是指一个模态来辅助或指导另一个模态工作. 互补性是指多模态数据利用各自的数据特点, 提供互补的信息进行工作. 根据联合学习的目标不同可以分为提升单模态任务性能和解决新的挑战性问题, 如图 7 所示. 虽然这两个目标的方向不同, 但是在具体任务中都突出了多模态数据相比于单模态数据的优势. 接下来, 本小节将分别介绍多模态数据在不同任务上的具体应用.

4.2.1 模态性能提升

多模态的联合学习有助于性能提升. 这类工作通过联合多种模态的信息来提高以往单模态任务上的性能, 比如, 视觉指导音频、音频指导视觉、深度指导视觉等.

视觉指导音频的任务有很多种, 比如, 视听语音识别和视觉指导的声源分离与定位等. 最早的自动语音识别系统几乎完全依赖声学语音信号, 无法适应含有噪声的环境 [154]. 视听语音识别引入了视觉信息, 极大地提高了噪声环境中的语音识别能力, 根据视觉信号补充受干扰的语音信息, 丰富语音内容. 考虑到视觉信息不受环境噪声的干扰, 研究者们从提取嘴唇轮廓和口腔区域强度中提取视觉特征来辅佐语音识别 [155], 性能优于仅用音频模态方法. 随着深度学习的加入, 研究者们利用神经网络来提取唇部特征并进一步提升了语音识别的准确度 [17]. 研究表明视觉信息有助于提高语音识别性能. 在音频中存在噪声的情况下, 视觉与音频模态的结合会带来显著的改善 [24]. 声源分离是音频处理中的经典问题, 仅靠音频模态进行声源分离则抛弃了视觉模块中的位置信息与发声物体的固有特性. 研

究者们利用机器去学习视觉场景和音频内容间的映射, 联合解析声音和图像, 实现视觉场景中的声源分离与定位 [47]. 演奏乐器的动作与乐器奏出的声音是相关联的, 因此研究者们根据演奏的运动线索提高分离乐器声音的性能 [156].

音频指导视频的任务包括视频摘要和视觉显著性预测等. 视频摘要不仅仅要考虑到视频帧的相似性, 还需要关注音频的变化 [157]. 研究者们利用音频相似性将视频分成几个场景, 联合视觉特征进行语义重要性计算, 提取关键帧 [48]. 在足球赛事中, 进球与扑救等精彩镜头都会伴随着突然变化的解说员的激动解说与现场球迷的震耳呼声, 这些音频都有助于视频摘要. 基于此, 研究者们提出了灵活高效的足球视频摘要系统 [158]. 为进一步提升视频摘要中音频与视频帧的联系, 研究者们考虑音频特征与视频特征的时间依赖性, 研究视频帧的全局依赖性 [159]. 融合视听信息来生成视频摘要在基准数据集上的实验表明, 基于音频的视频摘要方法优于仅用视觉信息的视频摘要方法 [157]. 视觉显著性预测是模仿人类的视觉特点来推测关键的区域, 但是人们通过视觉获取信息的同时都会伴随着音频的获取, 同时音频也会指导着人眼来观察感兴趣的位置. 为考虑被忽略的动态场景中无处不在的听觉信息, 研究者们提出一种新的针对丰富视听对应场景的多模态显著性模型 [160]. 鉴于之前的模型大多只针对于对话和人脸, 研究者们又提出了适用于任何场景的多模态显著性预测模型 [49]. 此外, 研究者们还提出包含 360° 的空间音频和视觉信息的视听显著性模型 [161].

深度指导视觉的任务有 RGB-D 目标检测与分割, 基于 RGB-D 的三维重建等. 深度模态提供的物体几何特征是视觉模态提供的色彩特征所不具备的. 因此很多研究者们把深度信息添加到目标检测与分割任务中, 辅佐 RGB 图像以提高性能. 研究者们利用深度信息对每个像素的离地高度和重力角度进行编码, 在性能上取得了较大的突破 [162]. 现有的 RGB-D 目标检测难点在于如何有效地跨模式融合, 研究者们首次利用三维卷积神经网络, 并在编码器和解码器两阶段进行初步和深度的跨模式融合 [163]. 由于现有方法利用深度图时, 大多集中在前景区域, 考虑到背景也提供了重要的信息, 研究者们引入双注意力机制模块, 来互补前景与后景的信息, 并取得了很好的性能 [164]. 基于 RGB-D 的三维重建与传统 SLAM [165] 方法主要区别在是否使用深度信息. RGD-D 传感器数据的噪声导致的几何误差的存在, 使得彩色图像无法精确对准重建的三维模型. 研究者们提出了一种从全局到局部的校正策略, 以获得更理想的三维重建 [166]. 为了解决 RGB-D 存在的数据失真问题, 研究者们设计算法, 以最好地利用 RGB-D 数据的优点, 同时抑制数据失真 [50].

4.2.2 模态创新应用

通过结合多种模态的信息可以解决以往单模态数据无法解决的问题, 完成单模态难以实现的任务, 例如复杂情感计算、音频匹配人脸建模、视听觉指导的音乐生成等.

传统的情感分析方法往往只能从单一模态的数据着手, 无法很好地获取有关语境的信息. 人类情感及语言表现形式的复杂性, 导致传统情感分析方法无法对复杂情感进行准确估计. 而基于多模态的情感分析方法除了提取单模态的特征外, 还会进行不同模态信息的融合得到更加全面的信息 [51], 因此会对语句的语境有着更加科学的估计, 从而实现对情感分析结果的正确估计 [167]. Pérez-Rosas 等 [52] 分别使用 OpenEAR 和 CERT 提取人物语音和面部表情的情感特征, 依据各种情感极性词在语句中出现的频率得到文本情感特征, 并将这 3 种特征利用特征级融合的方法进行融合, 送入 SVM 进行分析得到情感极性, 大大提高了准确率. Poria 等 [168] 则结合音频中音高和声音强度以及视频中连续图片的时间相关性, 利用一种卷积递归多核学习模型进行情感分析. 针对用语音特征区分愤怒和开心时准确率过低的问题, Hu 等 [169] 结合文本和语音更好地区分了愤怒和开心, 该方法用 openSMILE 提取声学特征, 采用基于词典的方法提取文本特征, 并进行特征级融合, 提高了愤怒与开心区分的准确率.

音频匹配人脸建模旨在通过图像与音频还原照片中人物说话的动态过程或赋予虚拟形象更加真实的对话体验 [53]. Xu 等 [54] 实现了虚拟模型嘴部动作对于音频的匹配. 近年来, Guo 等 [55] 依靠神经辐射场技术提出了一种由语音信号直接生成说话人视频的方法, 该方法仅需利用几分钟的视频, 即可利用任意音频模拟出视频中人物符合音频内容说话的场景. Prajwal 等 [170] 则专注于对嘴部动作的优化, 使得人物口型不再会因为动作变形, 解决了之前的方法中人物只能保持静止的问题. 此外, Müller-Eberstein 等 [171] 提出合成变分自编码器用于学习视觉信息和声音信息的协同, 进而实现基于视觉场景的配乐编曲. 此类任务也是多模态认知计算的未来发展方向之一.

4.3 分析与讨论

多模态协同在根本上是要协调不同的模态, 协同一致地完成某一目标. 在多模态协同过程中, 多模态数据利用模态融合增强模态间的相关性和互补性, 进而通过联合学习实现具体的任务. 因为多模态数据间的冗余性和异构性, 多模态数据在协同时需要着重考虑模态融合方法和联合学习方法. 对于模态融合来说, 融合可以发生在协同的不同时期并采用不同的融合方式. 因此, 为了达到理想的模态协同结果, 需要仔细分析模态各自的特点以及它们之间的相互关系来制定合适的融合时期和方式. 对于联合学习来说, 联合学习使得多模态数据得以组合并协同地完成一个任务. 根据任务的表现不同, 可以将联合学习分为提升性能和解决新任务. 实际上, 在联合学习中, 多模态数据往往不会共享一个网络, 然而在经过模态融合后模态数据会逐渐产生联系, 并最终产生协同效应.

研究者在构建模态融合和联合学习的基础上, 学习模态数据间的复杂联系, 实现了多模态或多模态间的交互协同. 本质上, 这种交互协同的产生来源于模态数据间固有的天然关系, 这种关系可以是兼容或互补的, 比如视听识别和声源分离. 有效地挖掘这种天然关系, 需要面向具体的模态数据和任务需求进行针对性分析, 如面向视频模态构建时空建模融合. 众多研究表明, 通过建立模态协同可以显著提升模型在相关单模态任务的表现, 甚至解决一些新的任务. 因此, 多模态协同是面向多模态应用场景的重要一环, 一些多模态应用出现和落地都依赖于模态数据间的有效协同.

5 多模态认知计算的难点和未来发展趋势

近年来, 深度学习技术在图像处理, 自然语言处理等领域取得了长足的发展, 推动着多模态认知计算向理论研究和工程任务的纵深发展. 在数据形式快速迭代和应用需求多元化发展的背景下, 多模态认知计算也面临新的问题和挑战. 从宏观角度来看, 前述所有任务都是围绕式 (5) 中的数据 (D)、信息量 (I)、融合机制 (A), 和任务 (T) 来提升机器认知能力 (ρ) 的. 本节将从以上 4 个方面对多模态认知计算当前的难点进行剖析, 并对未来的发展趋势进行展望和思考.

5.1 数据层面

在传统的多模态研究中, 数据的采集和计算通常是两个独立的过程. 例如, 在声源定位任务中, 分别利用麦克风阵列和摄像机采集场景中的音频和视频数据, 然后将数据传送到计算单元进行处理. 这种感算分离的处理方式是否存在缺点? 人类身处的世界是由连续模拟信号构成的, 而机器处理的是经过传感器时间采样和幅度量化后得到的离散数字信号. 从模拟信号到数字信号的转换过程, 必然造成信息的变形和丢失, 并加大计算负载. 20 世纪 70 年代以来, 微电子技术推动了数字化信号处理的持续发展. 而在未来, 信息将会以何种方式呈现给机器? 以光神经网络 (optical neural networks) 为代表的智能光电为这一难点提供了探索性的解决思路. 研究者利用非线性光学元件同时实现光信号的采集

和运算, 跳过了光信号到数字图像的转换步骤. Feldmann 等 [56] 更是提出了可用于监督学习和无监督学习的全光学神经突触网络, 并实现了光神经网络的可扩展. Mennel 等 [172] 的研究也呈现了类似的思路. 作者利用 WSe_2 半导体材料设计图像传感器阵列, 并利用静电掺杂技术使元件光敏性可调可控, 使传感器兼具感光和计算的功能. 这种光学传感器直接在光电模拟信号上进行计算, 有效地保留了原始信号中的信息. Xu 等 [173] 通过利用光频梳分离不同波长的光进行独立计算, 可以实现卷积核的并行, 显著提高了光神经网络的计算效率. 实际上, 光频梳是一个典型的时间同步工具, 在光原子钟、超级激光器、激光雷达等方面具有重要应用前景 [57,174]. 由于避免了与处理单元的数据传输, 数据转换和传送过程带来的能耗也大大减少. 在未来, 如果可以利用集成传感器完成多模态数据的感算一体, 实现模拟信号的多模态认知计算, 将会极大地提高机器的信息处理效率, 提升机器智能水平.

此外, 多模态数据获取不能仅仅停留在模仿人类感知上. 人类看不到的光谱波段, 听不到的声波频率, 都属于多模态认知计算的研究范畴. 除了视、听、嗅、味、触等人类知觉, 自然界还存在诸多感知世界的模态, 比如, 蝙蝠利用超声波识别障碍, 海豚通过脉冲进行交流等. 因此, 探寻潜在未知模态, 并挖掘其中包含的丰富信息, 将为多模态认知计算提供更广阔的发展空间.

5.2 信息层面

认知计算的关键是对信息中高级语义的表征、度量和评估, 包括一些客观的目标语义分析和主观的感受语义分析, 例如, 视觉中的位置关系、自然语言的起承转合、图像的风格、音乐的情感等. 人类相比于机器最显著的优势在于其具有更强的逻辑判断能力和情感认知能力, 这就是建立在人类具备的认知能力可以对信息中高级语义进行充分解析的基础之上的. 同时, 多模态认知计算涉及不同模态下高级语义信息的交互与理解. 目前多模态任务都局限于简单目标和场景下的交互, 一旦涉及更为深层的逻辑语义或主观语义就举步维艰. 例如, 机器可以生成一朵花开在草地上的图像, 但无法理解花草会在冬天凋谢的常识. 机器可以学习音乐的节奏快慢, 但无法判断它传递的情绪变化. 尽管符号主义人工智能可以依据专家知识描述特定的逻辑关系, 但在多模态任务中很难奏效, 因为不同模态对逻辑关系的表达差异很大. 同时, 具有强烈主观色彩的感受语义更难直接定义和度量, 差异化描述方式进一步加深了主观语义信息的度量难度. 因此, 如何去搭建不同模态下复杂逻辑和感受语义信息的通信桥梁, 建立特色的机器度量和评估体系是未来多模态认知计算的一大趋势.

5.3 融合机制层面

不同于单模态信息处理与建模, 多模态认知计算需要首先针对异质的模态信息进行高效的建模与分析, 进而实现高质量的关联、生成与协同. 对于差异化的模态信息 (如单通道音频的一维采样波形, 彩色图像的三通道像素等) 而言, 采用针对性的模型对不同模态进行建模是当前较为通用的方法. 然而, 为了能够实现多种模态信息同时输入下的高效认知, 多模态模型大多是由异构的单模态模型构成的. 当前, 多模态认知计算大多是在统一的学习目标下对模型进行优化的, 而这种优化策略缺乏对模型内部异构组成部分的针对性调整, 导致现有的多模态模型存在较大的欠优化问题 [175]. 如何对有异构部件组成的多模态模型进行高质量优化是当前融合机制层面的一个难点问题, 需要从多模态机器学习与优化理论方法等多方面切入, 实现多模态模型的高质量学习. 与此同时, 由于异构的单模态模型对于异质的模态信息建模能力不同, 在统一的学习目标下存在差异化的优化问题与优化进程, 这需要对异构子模型的优化进程进行监控, 实现多种模态信息在联合认知过程中的动态调控, 进而有效支撑高质量的关联、生成与协同等任务.

5.4 任务层面

在多模态认知计算中, 即便是同样的数据, 不同任务所关注的信息也有所区别, 这影响了机器的认知学习方式. 例如, 同样是 “听” 和 “看” 的任务, 视音频识别任务更关注音色, 而唇读任务则更关注语音内容. 如何设计任务反馈的学习策略, 同时提升多种相关任务的解决能力, 是多模态认知计算的难点之一. 人类在感受环境时, 利用多种感官与周围建立联系与互动, 并根据环境所给予的反馈形成具象, 深刻地综合感受. 而当前的机器学习则是从图像、文本等数据中感知和理解世界, 这种 “旁观式” 的学习方式难以建立对从数据到现实任务的真实映射. “具身智能” (embodied AI) [176] 为这一难题的解决提供了可能. 研究者提出, 智能体需要与外界环境进行交互, 在潜在模式中摸索解决问题的方法, 不断进化形成解决复杂任务的能力 [177]. 当前, 具身智能在视觉导航 [36]、语言学习 [178]、机器问答 [179] 等任务中得到了广泛研究. 如何使智能体与环境进行多模态交互, 形成 “具身认知”, 是未来的研究趋势之一.

此外, 随着虚拟现实和增强现实技术的发展, “元宇宙” 成为国内外各界共同关注的热门话题, 虚拟世界也从影视荧幕走向了大众的视线. 在未来, 虚拟世界无疑会产生大量多模态数据, 也将面临新的多模态计算任务, 这将为结合具身智能的多模态认知计算提供无限的发展动力和增长空间.

6 开放性问题讨论

6.1 人类认知与人工智能如何结合?

目前, 多模态认知计算的发展如火如荼. 大多数研究工作聚焦在人工智能领域, 致力于对视听嗅味触等多模态数据的分析, 以完成各种复杂任务. 在过去的几十年中, 人类的 “联觉”“知觉重塑” 和 “多通道知觉” 为多模态数据的关联、生成与融合提供了指导依据, 开启了多模态认知计算研究的序章. 但是, 人类认知存在太多未知和不确定. 人类认知是如何形成的? 其背后的机理是什么? 目前并不完全清楚. 缺乏认知进一步指导的多模态认知计算, 很容易陷入数据拟合的陷阱. 本文作者曾在视觉与学习青年学者研讨会 (VALSE) 上作为联合组织者发起过相关的线上 (2020) 和线下研讨会 (2022), 聚焦上述问题, 侧重从人类的多感官认知入手, 探究当前多模态相关研究与其的区别与联系.

未来, 多模态认知计算将如何迈向认知? 人类具有高可靠及较强泛化性能的多模态感知能力, 尤其是在部分感官能力缺失的情形下, 能够通过其他感官对缺失的能力进行一定补充. 认知神经科学家认为, 这种现象的潜在生理学基础可能是不同感官在信息编码中存在一个高级别语义的自组织关联网络, 该网络与特定模态类型无关, 但是可以直接关联到不同模态中, 从而实现高效的多模态感知. 对于多模态认知计算而言, 构建有效架构是提高多模态感知能力的关键一环. 本文认为, 可构建以 “元模态” 为核心的模态交互网络, 学习与特定模态类型无关的内在属性, 从而最大化关联与对齐不同的模态语义内容. 元模态指向一个紧致的低维空间, 可以实现到不同模态空间的投影, 从而具备更加泛化的表征能力.

6.2 多模态数据带来了什么?

近年来, 结合多模态数据的人工智能确实取得了更好的性能表现. 这显而易见, 在合理的模型优化方式下, 输入信息的增加往往会得到更好的结果. 但是, 再深入思考一下, 多模态数据到底带来了什么额外的信息, 又是如何提升性能的呢? 实际上, 多模态数据带来信息的同时, 也带来了大量噪声和冗余, 会出现信容降低的问题, 增加模型学习压力. 这会导致某些情况下, 多模态数据的性能不如单一

模态.

本文尝试从信息的角度给出如上问题的解释. 多模态信息之间具有相似性与互补性. 其中, 相似性部分是各个模态信息的交集, 即互信息, 代表了从不同模态描述同一场景的不同方面. 相似性部分对场景进行了更加综合的描述, 可以达到 “兼听则明” 的效果, 提升模型场景理解的鲁棒性. 互补性部分是各个模态信息的并集, 信息论里称为 “联合熵”, 代表了不同模态之间的差异性, 也包含噪声部分. 互补性部分是任一模态都不具备的, 它对单一模态的感知能力进行了拓展, 以获得更好的场景理解性能, 达到模拟人类联觉的能力.

6.3 多模态认知计算面临哪些真实场景?

现有多模态认知计算研究大都集中在图像视频数据中, 聚焦视听模态的分析. 这主要得益于近年来智能手机的普及和社交网络的快速发展, 使得图像视频数据爆炸式增长, 传播方式也日趋便利. 但是, 真实世界的多模态感知面临更加复杂的情况, 这里以机器人和临地安防为例.

机器人将是多模态认知计算的一个典型应用. 机器人的目的是像人类一样去感知去思考. 假设一个机器人要在真实环境中进行多模态感知, 首先, 要对视听嗅味触传感器进行集成, 目前针对前端传感器的研究明显不足. 然后, 感知要在三维空间中进行, 而不再是在视频画面中进行, 这就要求具有三维感知能力. 最后, 感知是在动态环境中进行的, 会存在机器与环境、各个模态与环境, 以及各个模态之间的交互, 这也是以后研究中需要重点考虑的.

临地安防 (vicinagearth security) 也为多模态认知计算提供了广阔的应用前景. 随着低空空域资源的逐渐释放和海洋开发能力的全面提升, 人工智能开始在涵盖低空、地上、水下的临地空间发挥作用, 涉及搜救、巡检等诸多安防问题. 以智能搜救为例, 无人机与地面无人设备的协同交互需要处理不同传感器产生的大量数据, 多模态认知计算也成为解决此类任务的关键核心技术之一, 需要与跨域智能交互、涉水光学等研究课题紧密结合. 同时, 临地安防对实时性和高效性的要求也对多模态认知计算提出了新的挑战. 在未来, 临地安防将成为多模态认知计算从理论走向应用的重要落地场景.

7 总结

信息领域的研究热点常常在获取 – 处理 – 反馈中迭代, 尤其前两者. 目前, 深度学习等处理方法的发展如火如荼, 下一个热点很可能是数据获取, 那么多模态将会迈入新的发展阶段. 本文抛砖引玉, 挂一漏万, 回顾了多模态认知计算的发展历程, 从理论、方法和趋势 3 个方面展开分析与思考. 首先, 构建信息传递模型刻画了机器从事件空间中提取信息的过程, 探讨了多模态认知计算的理论意义. 然后, 阐述了多模态关联、跨模态生成、多模态协同 3 个主线任务的理论联系, 对各项任务进行了统一. 通过对现有方法的分析与对比, 较为全面地展示了多模态认知计算的发展现状和关键技术. 进而, 结合当前人工智能的发展背景, 从信息度量、融合机制、学习任务和数据获取等方面探讨了多模态认知计算面临的挑战, 并讨论了未来值得探索的研究方向. 最后, 对多模态认知计算的开放性问题进行了一些设想. 实际上, 人类能感知到的模态信息是有限的. 人类仅可以看到 400~700 nm 的可见光, 这是光谱中很小的一部分; 只可以听到 20~20000 Hz 的可闻声波, 这也是声波很小的一部分. 庆幸的是, 借助各种先进的光电设备, 我们感知到了可见光和可闻声波之外的更多信息. 未来, 随着感知能力的进一步提升, 依托人类认知拓展物理感知边界, 实现信息域和认知域的统一, 是大势所趋. 希望本文能够为提升智能光电设备的感知能力和推动多模态认知计算的理论研究提供参考和启发.

参考文献

1 Li X L. Vicinagearth security. Commun CCF, 2022, 18: 44–52 [李学龙. 临地安防. 中国计算机学会通讯, 2022, 18: 44–52]
2 Gazzaniga M, Ivry R, Mangun G. Cognitive Neuroscience: The Biology of the Mind. New York: W. W. Norton & Company, 2002
3 Li X, Tao D, Maybank S J, et al. Visual music and musical vision. Neurocomputing, 2008, 71: 2023–2028
4 Cohen L G, Celnik P, Pascual-Leone A, et al. Functional relevance of cross-modal plasticity in blind humans. Nature, 1997, 389: 180–183
5 Zhang X, Li Z P, Zhou T, et al. Neural activities in V1 create a bottom-up saliency map. Neuron, 2012, 73: 183–192
6 Lin J, Men R, Yang A, et al. M6: a Chinese multimodal pretrainer. 2021. ArXiv:2103.00823
7 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
8 Rasiwasia N, Pereira J C, Coviello E, et al. A new approach to cross-modal multimedia retrieval. In: Proceedings of ACM International Conference on Multimedia, 2010. 251–260
9 Sharma A, Kumar A, Daume H, et al. Generalized multiview analysis: a discriminative latent space. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012. 2160–2167
10 Karpathy A, Joulin A, Fei-Fei L. Deep fragment embeddings for bidirectional image sentence mapping. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 1889–1897
11 Mansimov E, Parisotto E, Ba L J, et al. Generating images from captions with attention. In: Proceedings of International Conference on Learning Representations, 2016
12 Li X, Hu D, Lu X. Image2song: song retrieval via bridging image content and lyric words. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 5650–5659
13 Osman A, Samek W. DRAU: dual recurrent attention units for visual question answering. Comput Vision Image Understanding, 2019, 185: 24–30
14 Li X L, Zhao B. Video distillation. Sci Sin Inform, 2021, 51: 695–734 [李学龙, 赵斌. 视频萃取. 中国科学: 信息科学, 2021, 51: 695–734]
15 Morgado P, Li Y, Vasconcelos N. Learning representations from audio-visual spatial alignment. In: Proceedings of Advances in Neural Information Processing Systems, 2020
16 Chung J S, Zisserman A. Out of time: automated lip sync in the wild. In: Proceedings of Asian Conference on Computer Vision. Berlin: Springer, 2016. 251–263
17 Noda K, Yamaguchi Y, Nakadai K, et al. Audio-visual speech recognition using deep learning. Appl Intell, 2015, 42: 722–737
18 Wang J, Fang Z, Zhao H. AlignNet: a unifying approach to audio-visual alignment. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2020. 3309–3317
19 Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3674–3683
20 Parekh S, Essid S, Ozerov A, et al. Guiding audio source separation by video object information. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2017. 61–65
21 Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of International Conference on Machine Learning, 2015. 2048–2057
22 Bavelier D, Neville H J. Cross-modal plasticity: where and how? Nat Rev Neurosci, 2002, 3: 443–452
23 Cooke M, Barker J, Cunningham S, et al. An audio-visual corpus for speech perception and automatic speech recognition. J Acoust Soc Am, 2006, 120: 2421–2424
24 Afouras T, Chung J S, Zisserman A. LRS3-TED: a large-scale dataset for visual speech recognition. 2018. ArXiv:1809.00496
25 Cavazza M, Charles F. Dialogue generation in character-based interactive storytelling. In: Proceedings of Artificial Intelligence and Interactive Digital Entertainment Conference, 2005. 21–26
26 Hono Y, Hashimoto K, Oura K, et al. Sinsy: a deep neural network-based singing voice synthesis system. IEEE ACM Trans Audio Speech Lang Process, 2021, 29: 2803–2815

27 Yu C, Lu H, Hu N, et al. DurIAN: duration informed attention network for multimodal synthesis. 2019. ArXiv:1909.01700

28 Park T, Liu M, Wang T, et al. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 2337–2346

29 Deena S, Galata A. Speech-driven facial animation using a shared gaussian process latent variable model. In: Proceedings of International Symposium on Visual Computing, 2009. 89–100

30 Oh T, Dekel T, Kim C, et al. Speech2Face: learning the face behind a voice. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 7539–7548

31 Zhao B, Li X, Lu X, et al. Video captioning with tube features. In: Proceedings of International Joint Conference on Artificial Intelligence, 2018. 1177–1183

32 Zhao B, Li X, Lu X. CAM-RNN: co-attention model based RNN for video captioning. IEEE Trans Image Process, 2019, 28: 5552–5565

33 Yao B Z, Yang X, Lin L, et al. I2T: image parsing to text description. Proc IEEE, 2010, 98: 1485–1508

34 Yang Y, Teo C L, Daume H, et al. Corpus-guided sentence generation of natural images. In: Proceedings of Empirical Methods in Natural Language Processing, 2011. 444–454

35 Taylor S L, Kim T, Yue Y, et al. A deep learning approach for generalized speech animation. ACM Trans Graph, 2017, 36: 1–11

36 Liu X, Guo D, Liu H, et al. Multi-agent embodied visual semantic navigation with scene prior knowledge. IEEE Robot Autom Lett, 2022, 7: 3154–3161

37 Wu L, Xu M, Wang J, et al. Recall what you see continually using GridLSTM in image captioning. IEEE Trans Multimedia, 2020, 22: 808–818

38 Li X, Jiang S. Know more say less: image captioning based on scene graphs. IEEE Trans Multimedia, 2019, 21: 2117–2130

39 Cootes T F, Edwards G J, Taylor C J. Active appearance models. IEEE Trans Pattern Anal Machine Intell, 2001, 23: 681–685

40 Zhang R, Zhang Y, Lu C, et al. Unsupervised graph embedding via adaptive graph learning. IEEE Trans Pattern Anal Mach Intell, 2022. doi: 10.1109/TPAMI.2022.3202158

41 Freund Y, Schapire R. A short introduction to boosting. J Japan Soc Artif Intell, 1999, 14: 771–780

42 Fisher R A. The use of multiple measurements in taxonomic problems. Ann Eugenics, 1936, 7: 179–188

43 Bayoudh K, Knani R, Hamdaoui F, et al. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. Vis Comput, 2022, 38: 2939–2970

44 Liu Q, Wang W, Jackson P. A visual voice activity detection method with adaboosting. In: Proceedings of Sensor Signal Processing for Defence, 2011. 1–5

45 Shafer G. Dempster-Shafer theory. Encyclopedia Artif Intell, 1992, 1: 330–331

46 Katsaggelos A K, Bahaadini S, Molina R. Audiovisual fusion: challenges and new approaches. Proc IEEE, 2015, 103: 1635–1653

47 Zhao H, Gan C, Rouditchenko A, et al. The sound of pixels. In: Proceedings of European Conference on Computer Vision, 2018. 587–604

48 You J, Hannuksela M M, Gabbouj M. Semantic audiovisual analysis for video summarization. In: Proceedings of IEEE EUROCON 2009, 2009. 1358–1363

49 Tavakoli H R, Borji A, Kannala J, et al. Deep audio-visual saliency: baseline model and data. In: Proceedings of ACM Symposium on Eye Tracking Research and Applications, 2020. 1–5

50 Zollhöfer M, Stotko P, Görlitz A, et al. State of the art on 3D reconstruction with RGB-D cameras. Comput Graphics Forum, 2018, 37: 625–652

51 Zhang C, Yang Z, He X, et al. Multimodal intelligence: representation learning, information fusion, and applications. IEEE J Sel Top Signal Process, 2020, 14: 478–493

52 Pérez-Rosas V, Mihalcea R, Morency L. Utterance-level multimodal sentiment analysis. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, 2013. 973–982

53 Lahiri A, Kwatra V, Früh C, et al. LipSync3D: data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,

2021. 2755–2764
54 Xu Y, Feng A W, Marsella S, et al. A practical and configurable lip sync method for games. In: Proceedings of Motion in Games, 2013. 131–140
55 Guo Y, Chen K, Liang S, et al. AD-NeRF: audio driven neural radiance fields for talking head synthesis. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 5764–5774
56 Feldmann J, Youngblood N, Wright C D, et al. All-optical spiking neurosynaptic networks with self-learning capabilities. Nature, 2019, 569: 208–214
57 Feng Y, Xu X, Hu X, et al. Environmental-adaptability analysis of an all polarization-maintaining fiber-based optical frequency comb. Opt Express, 2015, 23: 17549–17559
58 Chen H, Xie W, Afouras T, et al. Localizing visual sounds the hard way. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021. 16867–16876
59 Hu D, Nie F, Li X. Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 9248–9257
60 Wu X, Wu Z, Ju L, et al. Binaural audio-visual localization. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 2961–2968
61 Sanguineti V, Morerio P, Bue A D, et al. Audio-visual localization by synthetic acoustic image generation. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 2523–2531
62 Qian R, Hu D, Dinkel H, et al. Multiple sound sources localization from coarse to fine. In: Proceedings of European Conference on Computer Vision, 2020. 292–308
63 Senocak A, Oh T H, Kim J, et al. Learning to localize sound sources in visual scenes: analysis and applications. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 1605–1619
64 Parida K K, Srivastava S, Sharma G. Beyond image to depth: improving depth prediction using echoes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021. 8268–8277
65 Hu D, Qian R, Jiang M, et al. Discriminative sounding objects localization via self-supervised audiovisual matching. In: Proceedings of Advances in Neural Information Processing Systems, 2020
66 Morgado P, Vasconcelos N, Langlois T R, et al. Self-supervised generation of spatial audio for 360° video. In: Proceedings of Advances in Neural Information Processing Systems, 2018. 360–370
67 Bredin H, Chollet G. Audiovisual speech synchrony measure: application to biometrics. EURASIP J Adv Signal Process, 2007, 2007: 070186
68 Rabiner L R, Juang B. Fundamentals of Speech Recognition. Upper Saddle River: Prentice Hall, 1993
69 Bertsekas D P. Dynamic Programming and Optimal Control. 3rd ed. Belmont: Athena Scientific, 2011
70 Aytar Y, Vondrick C, Torralba A. See, hear, and read: deep aligned representations. 2017. ArXiv:1706.00932
71 Monfort M, Jin S, Liu A, et al. Spoken moments: learning joint audio-visual representations from video descriptions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2021. 14871–14881
72 Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3156–3164
73 Antol S, Agrawal A, Lu J, et al. VQA: visual question answering. In: Proceedings of IEEE International Conference on Computer Vision, 2015. 2425–2433
74 Owens A, Efros A A. Audio-visual scene analysis with self-supervised multisensory features. 2018. ArXiv:1804.03641
75 Kazakos E, Nagrani A, Zisserman A, et al. EPIC-Fusion: audio-visual temporal binding for egocentric action recognition. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 5491–5500
76 Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 664–676
77 Li X L, Zhao Z Y. Pixel level semantic understanding: from classification to regression. Sci Sin Inform, 2021, 51: 521–564 [李学龙, 赵致远. 像素级语义理解: 从分类到回归. 中国科学: 信息科学, 2021, 51: 521–564]
78 Guhur P, Tapaswi M, Chen S, et al. Airbert: in-domain pretraining for vision-and-language navigation. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 1614–1623
79 McGurk H, MacDonald J. Hearing lips and seeing voices. Nature, 1976, 264: 746–748
80 Stiles N R B, Shimojo S. Auditory sensory substitution is intuitive and automatic with texture stimuli. Sci Rep, 2015, 5: 15628

81 Poirier C, de Volder A G, Scheiber C. What neuroimaging tells us about sensory substitution. Neurosci Biobehaval Rev, 2007, 31: 1064–1070

82 Striem-Amit E, Cohen L, Dehaene S, et al. Reading with sounds: sensory substitution selectively activates the visual word form area in the blind. Neuron, 2012, 76: 640–652

83 Assael Y M, Shillingford B, Whiteson S, et al. LipNet: end-to-end sentence-level lipreading. 2016. ArXiv:1611.01599

84 Cootes T F, Taylor C J, Cooper D H, et al. Active shape models-their training and application. Comput Vision Image Und, 1995, 61: 38–59

85 Matthews I, Cootes T F, Bangham J A, et al. Extraction of visual features for lipreading. IEEE Trans Pattern Anal Machine Intell, 2002, 24: 198–213

86 Owens A, Isola P, McDermott J H, et al. Visually indicated sounds. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2405–2413

87 Zhuang W, Wang C, Chai J, et al. Music2Dance: DanceNet for music-driven dance generation. ACM Trans Multimedia Comput Commun Appl, 2022, 18: 1–21

88 Hu D, Wang D, Li X, et al. Listen to the image. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7972–7981

89 Snoek C G M, Worring M. Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools Appl, 2005, 25: 5–35

90 Wang K, Yin Q, Wang W, et al. A comprehensive survey on cross-modal retrieval. 2016. ArXiv:1607.06215

91 Mignon A, Jurie F. CMML: a new metric learning approach for cross modal matching. In: Proceedings of Asian Conference on Computer Vision, 2012

92 Wang J, He Y, Kang C, et al. Image-text cross-modal retrieval via modality-specific feature learning. In: Proceedings of ACM on International Conference on Multimedia Retrieval, 2015. 347–354

93 Lu X, Wu F, Tang S, et al. A low rank structural large margin method for cross-modal ranking. In: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, 2013. 433–442

94 Wang W, Yang X, Ooi B C, et al. Effective deep learning-based multi-modal retrieval. VLDB J, 2016, 25: 79–101

95 Song G, Wang D, Tan X. Deep memory network for cross-modal retrieval. IEEE Trans Multimedia, 2019, 21: 1261–1275

96 Jiang Q, Li W. Deep cross-modal hashing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3270–3278

97 Hu D, Nie F, Li X. Deep binary reconstruction for cross-modal hashing. IEEE Trans Multimedia, 2019, 21: 973–985

98 Chen B, Rouditchenko A, Duarte K, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 7992–8001

99 Zhao B, Li X, Lu X. HSA-RNN: hierarchical structure-adaptive RNN for video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7405–7414

100 Vetter T, Poggio T. Linear object classes and image synthesis from a single example image. IEEE Trans Pattern Anal Machine Intell, 1997, 19: 733–742

101 Reed S E, Akata Z, Yan X, et al. Generative adversarial text to image synthesis. In: Proceedings of International Conference on Machine Learning, 2016. 1060–1069

102 Bailly G, Bérar M, Elisei F, et al. Audiovisual speech synthesis. Int J Speech Tech, 2003, 6: 331–346

103 Qiao T, Zhang J, Xu D, et al. MirrorGAN: learning text-to-image generation by redescription. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 1505–1514

104 Zen H, Tokuda K, Black A W. Statistical parametric speech synthesis. Speech Commun, 2009, 51: 1039–1064

105 Anderson R, Stenger B, Wan V, et al. Expressive visual text-to-speech using active appearance models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013. 3382–3389

106 Ren Y, Ruan Y, Tan X, et al. FastSpeech: fast, robust and controllable text to speech. In: Proceedings of Advances in Neural Information Processing Systems, 2019. 3165–3174

107 Wang W Y, Mehdad Y, Radev D R, et al. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016. 58–68

108 Zhu Z, Xu Z, You A, et al. Semantically multi-modal image synthesis. In: Proceedings of IEEE/CVF Conference on

Computer Vision and Pattern Recognition, 2020. 5466–5475
109 Schönfeld E, Sushko V, Zhang D, et al. You only need adversarial supervision for semantic image synthesis. In: Proceedings of International Conference on Learning Representations, 2021
110 Di X, Patel V M. Facial synthesis from visual attributes via sketch using multiscale generators. IEEE Trans Biom Behav Identity Sci, 2020, 2: 55–67
111 Elliott D, Keller F. Image description using visual dependency representations. In: Proceedings of Empirical Methods in Natural Language Processing, 2013. 1292–1302
112 Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proceedings of International Conference on Learning Representations, 2015
113 Devlin J, Cheng H, Fang H, et al. Language models for image captioning: the quirks and what works. In: Proceedings of Annual Meeting of the Association for Computational Linguistics, 2015. 100–105
114 Cho K, van Merrienboer B, Gülçehre Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of Empirical Methods in Natural Language Processing, 2014. 1724–1734
115 Liu S, Zhu Z, Ye N, et al. Improved image captioning via policy gradient optimization of spider. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 873–881
116 Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1179–1195
117 Xu N, Zhang H, Liu A A, et al. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. IEEE Trans Multimedia, 2020, 22: 1372–1383
118 Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. YouTube2Text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of IEEE International Conference on Computer Vision, 2013. 2712–2719
119 Che W, Fan X, Xiong R, et al. Visual relationship embedding network for image paragraph generation. IEEE Trans Multimedia, 2020, 22: 2307–2320
120 Wu J, Chen T, Wu H, et al. Fine-grained image captioning with global-local discriminative objective. IEEE Trans Multimedia, 2021, 23: 2413–2427
121 Chen L, Srivastava S, Duan Z, et al. Deep cross-modal audio-visual generation. In: Proceedings of Thematic Workshops of ACM Multimedia, 2017. 349–357
122 Zhou Y, Wang Z, Fang C, et al. Visual to sound: generating natural sound for videos in the wild. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3550–3558
123 Ghose S, Prevost J J. AutoFoley: artificial synthesis of synchronized sound tracks for silent videos with deep learning. IEEE Trans Multimedia, 2021, 23: 1895–1907
124 Hu D, Lu X, Li X. Multimodal learning via exploring deep semantic similarity. In: Proceedings of ACM Conference on Multimedia Conference, 2016. 342–346
125 Hu D, Wang C, Nie F, et al. Dense multimodal fusion for hierarchically joint representation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019. 3941–3945
126 Zhao B, Li H, Lu X, et al. Reconstructive sequence-graph network for video summarization. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 2793–2801
127 Zhang R, Guan L. Multimodal image retrieval via Bayesian information fusion. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2009. 830–833
128 Benmokhtar R, Huet B, Berrani S. Low-level feature fusion models for soccer scene classification. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2008. 1329–1332
129 Mangai U G, Samanta S, Das S, et al. A survey of decision fusion and feature fusion strategies for pattern classification. IETE Tech Rev, 2010, 27: 293–307
130 Wu C, Lee W, Chen Y, et al. Evolution-based hierarchical feature fusion for ultrasonic liver tissue characterization. IEEE J Biomed Health Inform, 2013, 17: 967–976
131 Nefian A V, Liang L, Pi X, et al. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP J Adv Signal Process, 2002, 2002: 783042
132 Li X L, Zhang H Y, Zhang R. Adaptive graph auto-encoder for general data clustering. IEEE Trans Pattern Anal

Mach Intell, 2022, 44: 9725–9732
133 Wang Q, Chen M, Nie F, et al. Detecting coherent groups in crowd scenes by multiview clustering. IEEE Trans Pattern Anal Mach Intell, 2020, 42: 46–58
134 Singh M, Singh S, Gupta S. An information fusion based method for liver classification using texture analysis of ultrasound images. Inf Fusion, 2014, 19: 91–96
135 Li X, Zhang H, Wang R, et al. Multiview clustering: a scalable and parameter-free bipartite graph fusion method. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 330–344
136 Snoek C, Worring M, Smeulders A W M. Early versus late fusion in semantic video analysis. In: Proceedings of ACM International Conference on Multimedia, 2005. 399–402
137 Atrey P K, Hossain M A, Saddik A E, et al. Multimodal fusion for multimedia analysis: a survey. Multimedia Syst, 2010, 16: 345–379
138 Lan Z, Bao L, Yu S I, et al. Multimedia classification and event detection using double fusion. Multimed Tools Appl, 2014, 71: 333–347
139 Bendjebbour A, Delignon Y, Fouque L, et al. Multisensor image segmentation using Dempster-Shafer fusion in Markov fields context. IEEE Trans Geosci Remote Sens, 2001, 39: 1789–1798
140 Xu H, Chua T S. Fusion of AV features and external information sources for event detection in team sports video. ACM Trans Multimedia Comput Commun Appl, 2006, 2: 44–67
141 Hu D, Li X, Lu X. Temporal multimodal learning in audiovisual speech recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 3574–3582
142 Gao R, Grauman K. 2.5D visual sound. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019
143 Yang K, Russell B, Salamon J. Telling left from right: learning spatial correspondence of sight and sound. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 9929–9938
144 Nojavanasghari B, Gopinath D, Koushik J, et al. Deep multimodal fusion for persuasiveness prediction. In: Proceedings of ACM International Conference on Multimodal Interaction, 2016. 284–288
145 Perez-Rua J, Vielzeuf V, Pateux S, et al. MFAS: multimodal fusion architecture search. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 6966–6975
146 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of International Conference on Learning Representations, 2015
147 Shih K J, Singh S, Hoiem D. Where to look: focus regions for visual question answering. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 4613–4621
148 Fan H, Zhou J. Stacked latent attention for multimodal reasoning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1072–1080
149 Sun C, Myers A, Vondrick C, et al. VideoBERT: a joint model for video and language representation learning. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 7463–7472
150 Gao J, Fan W, Jiang J, et al. Knowledge transfer via multiple model local structure mapping. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008. 283–291
151 Moon S, Kim S, Wang H. Multimodal transfer deep learning with applications in audio-visual recognition. 2014. ArXiv:1412.3121
152 Jin W, Sanjabi M, Nie S, et al. MSD: saliency-aware knowledge distillation for multimodal understanding. In: Proceedings of Findings of the Association for Computational Linguistics, 2021. 3557–3569
153 Agarwal D, Agrawal T, Ferrari L M, et al. From multimodal to unimodal attention in transformers using knowledge distillation. In: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance, 2021. 1–8
154 Yuhas B P, Goldstein M H, Sejnowski T J. Integration of acoustic and visual speech signals using neural networks. IEEE Commun Mag, 1989, 27: 65–71
155 Dupont S, Luettin J. Audio-visual speech modeling for continuous speech recognition. IEEE Trans Multimedia, 2000, 2: 141–151
156 Zhao H, Gan C, Ma W, et al. The sound of motions. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 1735–1744

157 Zhao B, Gong M, Li X. AudioVisual video summarization. IEEE Trans Neural Netw Learn Syst, 2021. doi: 10.1109/TNNLS.2021.3119969

158 Kiani V, Pourreza H R. Flexible soccer video summarization in compressed domain. In: Proceedings of IEEE International Conference on Computer and Knowledge Engineering, 2013. 213–218

159 Zhao B, Li X, Lu X. TTH-RNN: tensor-train hierarchical recurrent neural network for video summarization. IEEE Trans Ind Electron, 2020, 68: 3629–3637

160 Min X, Zhai G, Zhou J, et al. A multimodal saliency model for videos with high audio-visual correspondence. IEEE Trans Image Process, 2020, 29: 3805–3819

161 Chao F, Ozcinar C, Zhang L, et al. Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In: Proceedings of IEEE International Conference on Visual Communications and Image Processing, 2020. 355–358

162 Gupta S, Girshick R B, Arbeláez P A, et al. Learning rich features from RGB-D images for object detection and segmentation. In: Proceedings of European Conference on Computer Vision, 2014. 345–360

163 Chen Q, Liu Z, Zhang Y, et al. RGB-D salient object detection via 3D convolutional neural networks. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 1063–1071

164 Zhang Z, Lin Z, Xu J, et al. Bilateral attention network for RGB-D salient object detection. IEEE Trans Image Process, 2021, 30: 1949–1961

165 Davison A J, Reid I D, Molton N D, et al. MonoSLAM: real-time single camera SLAM. IEEE Trans Pattern Anal Mach Intell, 2007, 29: 1052–1067

166 Fu Y, Yan Q, Yang L, et al. Texture mapping for 3D reconstruction with RGB-D sensor. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4645–4653

167 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. 1097–1105

168 Poria S, Chaturvedi I, Cambria E, et al. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings of IEEE International Conference on Data Mining, 2016. 439–448

169 Hu T T, Chen L J, Feng Y Q, et al. Research on anger and happy misclassification in speech and text emotion recognition. Comput Technol Dev, 2018, 28: 124–127, 134

170 Prajwal K R, Mukhopadhyay R, Namboodiri V P, et al. A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of ACM International Conference on Multimedia, 2020. 484–492

171 Müller-Eberstein M, van Noord N. Translating visual art into music. In: Proceedings of IEEE/CVF International Conference on Computer Vision Workshops, 2019. 3117–3120

172 Mennel L, Symonowicz J, Wachter S, et al. Ultrafast machine vision with 2D material neural network image sensors. Nature, 2020, 579: 62–66

173 Xu X, Tan M, Corcoran B, et al. 11 TOPS photonic convolutional accelerator for optical neural networks. Nature, 2021, 589: 44–51

174 Wang W, Chu S T, Little B E, et al. Dual-pump Kerr micro-cavity optical frequency comb with varying FSR spacing. Sci Rep, 2016, 6: 28501

175 Peng X, Wei Y, Deng A, et al. Balanced multimodal learning via on-the-fly gradient modulation. 2022. ArXiv:2203.15332

176 Duan J, Yu S, Tan H L, et al. A survey of embodied AI: from simulators to research tasks. IEEE Trans Emerg Top Comput Intell, 2022, 6: 230–244

177 Gupta A, Savarese S, Ganguli S, et al. Embodied intelligence via learning and evolution. Nat Commun, 2021, 12: 5721

178 Özdemir O, Kerzel M, Wermter S. Embodied language learning with paired variational autoencoders. In: Proceedings of IEEE International Conference on Development and Learning, 2021. 1–6

179 Tan S, Xiang W, Liu H, et al. Multi-agent embodied question answering in interactive environments. In: Proceedings of European Conference on Computer Vision, 2020. 663–678

Multi-modal cognitive computing

Xuelong LI[1,2]

1. *School of Artificial Intelligence, Optics and Electronics* (*iOPEN*), *Northwestern Polytechnical University, Xi'an* 710072, *China*;
2. *Key Laboratory of Intelligent Interaction and Applications (Northwestern Polytechnical University), Ministry of Industry and Information Technology, Xi'an* 710072, *China*
E-mail: li@nwpu.edu.cn

Abstract The human brain perceives its surroundings through multiple sensory organs and integrates these multi-sensory perceptions to generate a comprehensive understanding. Inspired by synaesthesia, multi-modal cognitive computing endows machines with multi-sensory capabilities and has become the key to general artificial intelligence. With the explosion of multi-modal data such as image, video, text, and audio, a large number of methods have been developed to address this topic. However, the theoretical basis of multi-modal cognitive computing is still unclear. From the perspective of information theory, this paper establishes an information transmission model to profile the cognitive process. Based on the theory of information capacity, this study finds out that multi-modal cognitive computing helps machines extract more information. In this way, multi-modal cognitive computing research is unified by the same theoretical basis. Then, the development of typical tasks is reviewed and discussed, including multi-modal correlation, cross-modal generation, and multi-modal collaboration. Finally, focusing on the opportunities and challenges faced by multi-modal cognitive computing, some potential directions are discussed in depth, and several open-ended questions are considered.

Keywords artificial intelligence, multi-modal, cognitive computing, synaesthesia, information capacity