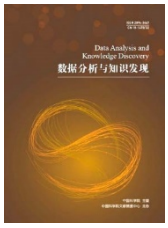


数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：　多模态命名实体识别研究进展
作者：　韩普，陈文祺
网络首发日期：　2023-08-29
引用格式：　韩普，陈文祺．多模态命名实体识别研究进展[J/OL]．数据分析与知识发现.
https://link.cnki.net/urlid/10.1478.G2.20230829.0924.002

网络首发时间：2023-08-29 14:21:23
网络首发地址：https://link.cnki.net/urlid/10.1478.G2.20230829.0924.002

多模态命名实体识别研究进展

韩　普[1,2]，陈文祺[1]
[1](南京邮电大学管理学院 南京 210003）
[2](江苏省数据工程与知识服务重点实验室 南京 210023）

摘要：【目的】梳理归纳多模态命名实体识别研究成果，为后续相关研究提供参考与借鉴。【文献范围】以 Web of Science、IEEE Xplore、ACM digital library、知网数据库为检索来源，以多模态命名实体识别、多模态信息抽取以及多模态知识图谱为检索词进行文献检索，共筛选出 83 篇代表性文献。【方法】从概念、特征表示、融合策略和预训练模型四个方面对多模态命名实体识别研究进行论述总结，指出现存问题和未来研究方向。【结果】多模态命名实体识别目前主要围绕模态特征表示和融合两方面展开且在社交媒体领域取得了一定进展，需要进一步改进多模态细粒度特征提取和语义关联映射方法以提升模型的泛化性和可解释性。【局限】直接以多模态命名实体识别为研究主题的文献数量较少，在支撑综述结果方面存在局限性。【结论】针对多模态命名实体识别亟需解决的问题展望未来发展趋势，为进一步拓宽多模态学习在下游任务应用的研究范畴，破解模态壁垒和语义鸿沟提供了新思路。

关键词：多模态实体识别；特征表示；多模态融合；多模态预训练
分类号：TP391

Review of Research Progresses in Multimodal Named Entity Recognition

Han Pu[1,2], Chen Wenqi[1]
[1] (School of Management, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)
[2] (Jiangsu Provincial Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China)

Abstract: **[Objective]** The research results of multimodal named entity recognition are sorted out and summarized to provide reference and reference for subsequent related research. **[Coverage]** Using Web of Science, IEEE Xplore, ACM digital library, CNKI databases as the search sources, 83 representative papers were selected by using multimodal named entity recognition, multimodal information extraction and multimodal knowledge graph as the search terms. **[Methods]** We summarize the multimodal named entity recognition research in four aspects: concepts, feature representation, fusion strategies and pre-trained models, and propose existing problems and future research directions. **[Results]** Multimodal named entity recognition is currently focused on both modal feature representation and fusion and has made some progress in the field of social media. Further improvements in multimodal fine-grained feature extraction and semantic association mapping methods are needed to enhance the generalization and interpretability of the model. **[Limitations]** The small amount of literature that directly focuses on multimodal named entity recognition as a research topic has limitations in supporting the results of the review. **[Conclusions]** In view of the urgent problems of multimodal named entity recognition looking into the future development trend, it provides

new ideas to further broaden the research scope of multimodal learning in downstream task applications and to break the modal barriers and semantic gaps.
Keywords: Multimodal named entity recognition; Feature representation; Modal fusion; Multimodal pre-training

1 引言

近年来，随着互联网的飞速发展，网络信息正逐渐从文本、图像、音频、视频等单模态形式过渡到相互融合的多模态形式，这些多模态数据包含了丰富的实体形态信息和语义关联知识，围绕多模态展开的相关研究受到多个学科和领域的极大关注。多模态命名实体识别（Multimodal Named Entity Recognition，MNER）旨在通过融合文本、图像、声音和视频等多种模态，实现语义信息在各个模态之间的交流和转换，进而准确识别目标模态所包含的实体[1-3]。多模态命名实体识别是自然语言处理、计算机图形学、多媒体处理和语音识别等学科领域的交叉研究，是跨模态信息检索[4]、多模态机器翻译[5]、视觉问答[6]和多模态知识图谱构建[7]等多模态学习任务的基础环节，同时也是多模态学习打破不同模态语义鸿沟，获得更强的语义理解、知识补全和知识推理能力的重要突破点。

多模态命名实体识别是多模态学习和自然语言处理领域中一项重要的基础研究，目前相关研究还处在探索阶段，为系统了解多模态命名实体识别最新的研究进展，本文在 Web of Science、IEEE Xplore 和 ACM digital library 数据库以“TS=（Multimodal Named Entity Recognition OR Multimodal Information Extraction OR Multimodal Knowledge Graph）”为检索式进行检索，在中国知网以“(SU = '命名实体识别' OR SU = '实体识别') AND (SU='多模态' OR SU ='跨模态') OR SU='多模态知识图谱' OR SU='多模态信息抽取'”为检索式进行检索。考虑到多模态命名实体识别是近十年来引起学界广泛关注的研究主题，本研究将检索文献时间范围设为 2010 年至 2023 年，另外通过回溯检索，最终人工筛选出 83 篇与主题密切相关的学术文献，其中英文文献 73 篇，中文文献 10 篇。

通过梳理相关文献发现，目前多模态命名实体识别相关的综述文献主要围绕多模态机器学习技术和算法展开。其中较为经典的是卡内基梅隆大学研究团队发表的两篇综述[8,9]，两篇文献详细阐述了多模态机器学习算法的发展现状、面临挑战和发展趋势。国内也有学者对多模态机器学习应用、多模态融合和多模态预训练等技术做了归纳总结[3,10,11]。尽管目前已有不少针对多模态命名实体识别的深度学习模型，但是学界对多模态命名实体识别尚未形成成熟的研究思路和研究框架，对于多模态命名实体识别任务涉及的模态表示、模态融合和预训练等关键技术还缺乏系统的梳理。基于此，本文首先对多模态命名实体识别的概念与研究框架进行论述，接着阐述了多模态学习在实体识别任务中的关键环节，重点对模态特征表示和多模态数据融合的关键技术进行深入分析，并归纳了多模态预训练在多模态命名实体识别任务中的最新研究进展，最后对已有研究遇到的问题和挑战进行总结和展望。

2 多模态命名实体识别概述

多模态（Multimodality）是指事物呈现的不同方式[3,8]。目前大量网络信息以图文、声音和视频等多模态形式出现，尤其是近些年涌现出大量短视频、直播和视频会议等应用生成了海量的多模态数据，这为多模态学习提供了丰富的数据资源和应用场景。多模态命名实体识别是在多模态机器学习研究范式下，以传统的文本实体为基础，引入图像或声音等不同模态数据强化机器对不同模态信息的感知、理解、推理和学习。作为多模态信息抽取的关键环节，近些年多模态命名实体识别受到学界的格外关注。本节将对多模态命名实体识别的发展历程、概念和任务以及研究框架进行分析。

2.1 多模态命名实体识别发展历程

多模态命名实体识别任务最早由 Moon 等[1]提出，源于实体抽取任务在社交媒体领域中的应用，旨在提升实体识别的准确率。为了更清晰呈现多模态命名实体识别发展脉络，本文将该任务划分为模态引入、模型改进和领域推广三个阶段。在模态引入阶段，针对社交平台上推文在文本基础上增加了图片模态的变化，命名实体识别任务尝试利用视觉特征补充文本上下文信息以提升实体识别效果，并在此基础上衍生出了多模态命名实体识别任务。该阶段研究主要基于多模态属性抽取任务框架确立了多模态命名实体识别的研究思路[12,13]，并构建了面向该任务的数据集。随着多模态命名实体识别研究的推进，早期工作出现了图片和文本内容不匹配无法实现语义关联的问题，且图片中与文本无关的信息可能作为噪音而影响模型性能。针对上述问题，在多模态命名实体识别模型改进阶段，现有研究从不同视角提出了多种思路和方法。在多模态语义对齐方面，Yu 等[13-15]提出基于 Transformer 架构的多模态命名实体识别模型，该模型利用 Transformer 强大的语义捕捉能力学习图片与文本实体相关联的语义特征；Zhang 等[16]和 Zheng 等[17]分别提出了基于多模态图和对抗性双线性注意力融合方法提取细粒度语义特征以实现语义关联。在缓解视觉噪声方面，Sun 等[18,19]通过多模态预训练模型和关系传播机制实现了视觉噪声过滤，进而提升模型识别实体效果。从现有研究可以发现，目前针对社交媒体领域的多模态命名实体识别任务已比较成熟，有部分研究尝试在多个领域拓展该任务的研究范畴。在领域推广阶段，多模态命名实体识别研究主要聚焦在模态拓展和领域迁移方面，其中模态拓展是指研究对象不仅针对图像和文本，还进一步拓展到音频、视频等模态[20,21]；而领域迁移是指数据来源不局限于社交平台，还逐渐扩展到医疗健康、数字人文和语言文学等领域[3,22,23]。

2.2 多模态命名实体识别任务概念和分类

目前学界对多模态命名实体识别尚未形成统一认可的概念，已有研究主要从不同视角给出了多模态命名实体识别的概念和解释，Moon 等[1]将多模态引入实体识别任务以补充高噪声短文本的语义信息；Zhang 等[2]指出多模态命名实体识别主要是借助图像等多模态抽取文本中的实体，并按照预定义类型对识别出的实体进行分类；Lu 等[12]认为多模态命名实体识别任务是利用图像辅助文本实体识别，该任务本质上仍为序列标注问题；Yu 等[14]提出多模态命名实体识别是利用关联图像以更好地识别文本中的实体；范涛等[22]认为多模态命名实体识别是为了挖掘文本和图片的关系以增强文本语义信息，从而提升模型识别实体的性能。通过已有研究分析可知，多模态命名实体识别是在多模态学习研究框架下传统命名实体识别任务的外延，其主要目的是通过融合多种互为补充的模态信息，提高实体识别的准确性、鲁棒性和泛化性，为多模态语义深层交互和理解奠定基础。尽管学界针对多模态命名实体识别尚未达成统一认可的概念，但对于该任务试图达到的目标是一致的，即在语言理解和视觉环境之间架起桥梁，增强机器对信息和知识的理解。

根据任务设计的模态数据组合划分，多模态命名实体识别可分为基于图像—文本的命名实体识别和基于声音—文本的命名实体识别两类，前者最初聚焦于社交媒体领域以解决高噪声短文本中的实体识别问题，并逐步迁移至其它领域以补充文本的语义信息[24-26]，也是目前多模态命名实体识别研究中最为广泛的一类任务；后者主要利用声学模态中语调、节奏等语义信息提升文本模态实体识别效果，有助于确定文本中实体边界和消除实体歧义问题[21,27]。

2.3 多模态命名实体识别研究框架

目前常见的主流多模态实体识别框架通常是在传统单模态模型基础上融入多模态机器学习方法。具体而言，多模态命名实体识别遵循特征表示、特征学习和解码预测的研究思路，仍然属于序列标注任务。但与基于文本的单模态命名实体识别任务相比，多模态命名实体识别研究

在单独提取不同模态特征基础上需要进一步实现模态间的语义对齐和融合，即构建能够融合异构数据的共享语义子空间以提升模型对多模态数据的语义理解，最终提高实体识别效果[9,28,29]。图 1 给出了目前最为常见的图像—文本多模态实体识别的研究框架。该示例中，首先将输入的单模态数据分别编码为向量形式，然后输入多模态交互模块中执行对齐和融合任务，通过特征学习构建多模态表示，最后解码输出实体标签[30,31]。

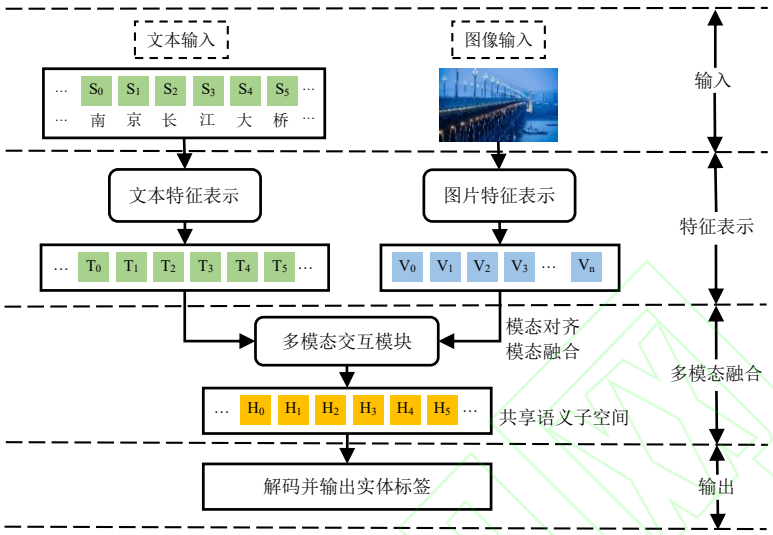


图 1 多模态命名实体识别研究框架（以文本-图像为例）

Fig 1. A Research Framework for Multimodal Named Entity Recognition (With Text-Image Example)

3 模态的特征表示

在多模态命名实体识别任务中，模态特征表示是多模态融合的前提，是构建共享语义子空间的基石。模态特征表示旨在通过神经网络模型对不同模态数据进行线性或非线性映射，以生成模态的高阶语义特征表示并以特征向量形式呈现[32,33]。目前在多模态实体识别研究中模态特征表示的研究重点集中在模态更深层次、更细粒度的语义信息提取。本节重点对多模态命名实体识别任务中主要涉及的文本、图像和声音三种模态特征表示方法进行分析。

3.1 文本特征表示

文本特征表示的核心是对文本的基本语义单元进行表示[8]。目前多模态实体识别处于以文本模态实体为主要识别对象，引入图像、声音等模态作为辅助以提升模型识别精度的阶段，文本特征表示依然是多模态命名实体识别任务的研究热点。在多模态命名实体识别任务中常见的文本特征表示模型有卷积神经网络（CNN）、双向长短期记忆网络（BiLSTM）和 BERT 等[34-36]。其中，CNN 通过捕捉局部特征实现对文本实体的特征表示和学习；BiLSTM 在循环神经网络基础上增加了门控机制和记忆单元，能够更好地捕捉长距离的双向语义依赖，有效地解决了特征学习中长程依赖性问题；BERT 使用海量无监督数据进行预训练，进而预先学习大量的语义知识，并利用微调技术适应多种下游任务，具有极强的语义表征优势和领域适应性。总体而言，文本模态表示越来越注重细粒度特征的学习，通过深层语义信息的充分挖掘提升机器对自然语言的感知能力。

3.2 图像特征表示

图像特征表示是多模态命名实体识别任务的研究重点，从图像中捕获与实体相关的高层语义信息并融合文本特征以提高性能是多模态实体识别任务的核心[3,8,37]。随着计算机视觉技术的发展，多模态实体识别任务图像特征表示出现了全局图像特征、局部图像特征和视觉对象特征

表示等不同语义密度的特征表示方法[37]。全局图像特征是将整张图片编码为一个静态向量，并通过神经网络提取与文本内容相关的视觉信息，该方法简单易操作但可能会导致大量图像细节信息的丢失[38-40]。局部图像特征表示通过将整张图片平均分为多个视觉区域，然后显式建模文本序列与视觉区域之间的相关性，最后得到图像的特征矩阵[41,42]。相较于全局图像特征提取方法，局部图像特征提取尽可能保留了图像的细粒度特征信息，但可能导致模态对齐困难和计算资源消耗较大等问题。视觉对象特征表示在局部图像特征表示的基础上实现了更细粒度的视觉特征表示。常见的 Fast RCNN 等目标检测模型以边界框形式标注图像中的实体并生成弱文本标签，然后进一步结合注意力机制实现模态对齐[43]。在多模态命名实体识别任务中，视觉对象表示能够减少图像噪声来提升图像与文本的语义匹配程度[28]，不足之处是难以刻画图像中多个实体目标之间的语义联系。

3.3 声音特征表示

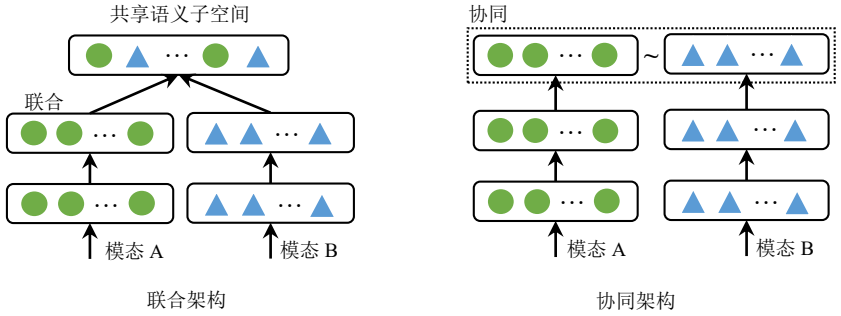
声音模态特征表示是指通过提取声音信号的语义特征向量，利用神经网络模型将原始声音信号映射到一个连续的向量空间。构建声音特征表示的模型主要由语音处理层和编码器—解码器结构两个部分组成[23,44-46]。语音处理层主要目标是将波形转换为向量序列，通常使用梅尔倒谱系数（Mel-scale Frequency Cepstral Coefficients，MFCC）作为特征向量，MFCC 是能够准确描述声道形状在语音短时功率谱上的一种声学特征[44]。声音特征向量经过神经网络多级映射后得到高阶特征表示，能够学习音频数据中多层次语义信息[45,46]。目前针对文本—声音模态的命名实体识别研究成果还比较少，但是声音特征中包含节奏、情感、音调和压力等语义信息有助于模型明确实体边界和解决实体多层嵌套问题[21,27]，因此融入声音特征的多模态命名实体识别模型有较大提升空间。

4 多模态实体识别中的多模态融合策略

多模态融合（Multimodal Fusion）是多模态命名实体识别任务中一个极具挑战性的问题，它是指将具有异质性的多种单模态特征映射至同一向量空间中，并对该空间中的多模态信息进行处理，最终构建多模态特征向量表示[8,9]。多模态融合是体现多模态优势的核心技术，通过多种模态交互并构建融合多种异构信息的特征表示使得模型能够持续的做自适应调整，从而提高模型的精确性和稳定性[9]。本文将从多模态融合架构和融合方法详细分析多模态融合在多模态命名实体任务中的应用。

4.1 多模态融合架构

根据多模态特征表示方式，多模态命名实体识别中的多模态融合架构可以分为联合架构和协同架构[8]，示意图如图 2。联合架构是将不同模态的特征表示融合到一个共享语义子空间，通过对齐单模态特征向量得到多模态联合特征表示[22,47,48]；协同架构旨在保证单模态特征相对独立的前提下为其它模态提供语义信息补充，最终生成多模态协同特征表示[4,13,49]。



联合架构　　协同架构

图 2 两种多模态融合架构示意图
Fig 2. Schematic Diagram of Two Multimodal Convergence Architectures

（1）联合架构

联合架构是将单模态特征表示映射到多模态共享语义子空间，从而构建融合多个模态语义信息的多模态特征向量。联合架构在多模态分类和预测任务中使用广泛并且拥有较好表现，如视频分类[50]、命名实体识别[51]、情感分析[52]、视觉问答[53]和语音识别等[54]。联合架构实现的关键在于构建能够统一表示多种语义信息的向量空间，早期工作一般采取简单拼接方式[8,9]，即在不同的隐藏层实现共享语义子空间，将转换后的各个单模态特征向量语义组合在一起，如式（1）所示。另一种方法是将多种模态融合在统一的张量中，该张量由所有单模态特征向量的输出乘积构成[54]，具体如式（2）。

$$y = f_1(x_1) + f_2(x_2) \quad (1)$$

$$y = f_1(x_1) \times f_2(x_2) \quad (2)$$

其中，x_i 表示输入模态，f_i 表示模态 x_i 的编码方式，y 表示融合后的多模态特征表示。

简单拼接方法虽然操作简便，但是容易造成语义丢失问题从而导致模型性能下降，因此目前常借助深度神经网络更加充分的融合模态特征[55]，深度神经网络通过端到端的训练模式保证了多模态融合过程中语义的连续，在构建多模态联合特征表示的同时也可以执行回归或预测等具体任务。

（2）协同架构

多模态协同架构是将多种单模态特征表示在一定约束条件下实现协同工作[56,57]。协同架构在保持各个单模态特征相对独立的同时实现多模态融合，这一特性有利于在跨模态转移学习中实现不同模态或领域间信息的相互传递[58]。具体地，在多模态命名实体识别任务中协同架构常用于基于跨模态转化的融合方法，一般思路是首先分别学习文本、图像等单模态特征，然后将图片等其它模态信息在一定约束条件下转化为文本信息，最后将转化后的文本辅助信息与原文本信息结合从而实现多模态融合[4,28,59]。尽管协同架构能够避免在同一语义空间中异构数据的语义冲突问题，但是该融合思路对构建高效的跨模态学习模型有较高的要求，因此协同架构在多模态实体识别任务中应用较少。

4.2 多模态融合方法

早期的多模态融合方法研究主要集中在早期融合和晚期融合，这两种方法也被称为与模型无关的融合方法。早期融合方法通过简单拼接将提取的单模态特征表示融合为多模态特征表示[20,60,61]，该方法操作简单容易实现，但无法充分利用多个模态间语义的关联性和互补性，且存在引入噪音信息的不足[63,63]。晚期融合则针对模态间差异训练不同的模型，然后再融合多模态语义特征并构建多模态特征表示。晚期融合方法相比早期融合方法鲁棒性更强，但是由于融合过程与模态特征无关，因此难以实现多模态间的交互[64,65]。为实现多模态异构信息更有效的交互，目前常用方法是基于模型的融合，该方法是在神经网络模型基础上，通过对齐、转换等操作将各单模态特征映射至统一的语义向量空间，构建多模态特征表示。鉴于模型融合方法逐渐成为当前的研究热点，本小节将重点介绍在多模态命名实体识别任务中该方法的具体应用。

（1）注意力融合

① 视觉注意力融合

视觉注意力融合最早由 Lu 等[12]运用在多模态命名实体识别任务中，该融合方法旨在找出图

片中与文本密切相关的区域，同时过滤与上下文不相关的视觉噪音信息。在视觉注意力模块中，首先将文本特征作为查询向量，把局部视觉特征作为键和值，通过感知机将文本特征和视觉特征投影到同一维度；然后通过点积实现视觉注意力机制，并使用门控机制动态控制视觉特征与文本特征融合；最后将融合后的特征向量输入至 CRF 层进行标记输出，视觉注意力机制具体见式（3）~（5）。

$$A=\left(tanh(W_tQ)\right)\oplus\left(tanh(W_vV)\right) \tag{3}$$

$$E=softmax(W_aA+b_a) \tag{4}$$

$$v_c=\sum a_iv_i \quad a_i\in E,\ v_i\in V \tag{5}$$

其中，Q 和 V 分别为文本和视觉的特征向量，E 为局部视觉特征的权重，v_c 为视觉上下文特征向量，该向量经过门控机制过滤噪声后即可得到多模态特征表示[12]。

视觉注意力机制能够将模型集中关注的图像区域可视化，直观地展现了文本信息和图像信息在语义空间中是否对齐，也有利于增强多模态融合模型的可解释性。

② 协同注意力融合

协同注意力融合是由 Zhang 等[2]首次提出并成功应用与于多模态命名实体识别任务，该方法通过协同注意力机制模块实现多模态融合。基于协同注意力机制的融合模型包括文本引导视觉注意力、图像引导文本注意力和门控机制三部分。文本引导图像注意力指计算文本中给定词与图片中各区域的相关程度；而图片引导文本注意力指计算给定图像与文本的相关程度，通过图像捕获文本内部的语义联系。文本和图像依次引导注意力机制旨在对齐图像和本文特征并过滤视觉噪音信息，以提升模型的特征表达能力。门控机制主要由融合门和过滤门组成，主要目的是获取更高质量的多模态特征表示，最后将该多模态特征向量输入至 CRF 层进行解码并获取标签[66,67]，具体见式（6）~（8）。

$$\hat{t}_i=WGA(\theta_w;m_t,t_i) \tag{6}$$

$$\hat{m}_t=IGA(\theta_i;\hat{t}_i,m_t) \tag{7}$$

$$u_t=m_t\oplus gate(\hat{t}_i,\hat{m}_t) \tag{8}$$

其中，WGA 表示文本引导图像注意力，IGA 表示图像引导文本注意力，t_i 表示图片特征向量，m_t 为文本特征向量[22]。

③ Transformer 注意力融合

Transformer 注意力融合由 Yu 等[14]提出并较早应用于多模态命名实体识别任务，该融合方法在 Transformer 模型基础上增加了多模态交互模块，用于捕捉文本和图像之间的语义联系。具体地，首先通过 BERT 模型和 ResNet 模型分别得到文本和图像的特征表示；接着将两种模态的特征向量输入至多模态 Transformer 模块以实现不同模态的交互，并利用跨模态 Transformer 层和视觉门控对齐文本与图像特征；最后将构建的多模态特征表示置入条件随机场融合解码得到实体标签[12,18]，多模态交互模块是 Transformer 注意力融合方法的核心，多模态交互的原理见式（9）~（10）。

$$A=f_M(IAW(P),WAI(Q)) \tag{9}$$

$$g=sigmoid(W,A)\cdot Q \tag{10}$$

其中，P 和 Q 分别为文本和图像特征，IAW 表示视觉引导文本特征表示，WIA 表示文本引导视觉特征表示，A 为视觉和文本互相引导后经门控机制过滤后得到的特征向量[14]。

④ 双线性注意力融合

双线性注意力模型是双线性池化的扩展，增强了模态间的交互。为解决注意力机制应用于多模态学习任务中耗费过多的计算资源问题，Kim 等[68]提出了双线性注意力网络（Bilinear Attention Networks，BAN），使用低秩双线性池化层提取多模态输入的联合表示，在不增加成本的情况下高效地实现了模态间交互。为了更好地捕捉视觉对象和文本实体之间的关系，Zheng 等[17]在多模态命名实体识别任务中，利用门控机制改进了双线性注意力神经网络。改进后的网络能够通过计算输入通道匹配对之间的相关性，学习到不同模态之间的交互联系，门控机制是为了过滤无关的视觉噪音，最后把文本特征和视觉特征相加得到多模态特征表示，并输入到 CRF 层进行解码。双线性注意力融合机制的原理见式（11）~（13）。

$$f = BAN\left(X, Y, A\right) \tag{11}$$

$$A = softmax\left(\left(\left(I \cdot p^{T}\right) \circ X^{T}U\right)V^{T}Y\right) \tag{12}$$

$$Y_{att} = \left(U_{\alpha}Y\right)A^{T} \tag{13}$$

其中，X 和 Y 分别表示文本和图像两种模态通道的输入，A 代表双线性注意力图，I 为单位矩阵，p 为池化参数矩阵，U、V 和 U_{α} 为参数矩阵，$\circ$ 表示 *Hadamard* 乘积，Y_{att} 是句子中每个字所对应的视觉特征[17,68]。

（2）图模型融合

基于注意力机制融合的方法往往难以捕捉到图文对中细粒度语义单元之间的关系。为了更充分利用视觉信息，有学者提出了一种基于图的多模态融合方法[19,69]，该方法首先构建多模态图表示文本和图像的特征输入，其中每个节点表示一个单词或视觉目标，每条边表示模态间或模态内的联系；然后将多模态图送入多层感知机依次更新所有节点状态并编码生成多模态特征表示；最后通过 CRF 解码器输出实体标签。多模态图融合的计算公式如式（14）~（17）。

$$C_{x,o}^{(l)} = multihead(H_{x,o}^{(l-1)}, H_{x,o}^{(l-1)}, H_{x,o}^{(l-1)}) \tag{14}$$

$$R_{x_i,o_j}^{(l)} = C_{x_i,o_j}^{(l)} + \sum a_{i,j} \otimes C_{o_j,x_i}^{(l)} \tag{15}$$

$$H_{x}^{(l)} = FFN(R_{x}^{(l)}) \tag{16}$$

$$H_{o}^{(l)} = FFN(R_{o}^{(l)}) \tag{17}$$

其中，H_x 表示文本节点，H_o 表示图像节点，R_x 和 R_o 分别表示文本和图像融合上下文信息后特征表示，FFN 为前馈神经网络[16]。

基于多模态图的融合模型表现出较强的关系推理能力，因此可以更准确地捕捉多模态之间的语义联系，同时也能够处理更加复杂的异构数据。

（3）跨模态检索融合

跨模态检索融合旨在实现不同模态之间的信息转换，主要研究思路是通过一种模态样本来检索具有近似语义的另一种模态样本[70,71]。Yao 等[4]提出使用跨模态密集检索的方法来融合各模态特征，该方法借助 Vokenization 视觉语言模型将文本和图像的特征表示映射到一个共享语义子空间，通过最大内积搜索匹配文本和图像特征进而实现多模态融合，该融合方法具体过程见式（18）~（20）。

$$f_{\theta}(w_i; s) = Encoder_{T}(s) \tag{18}$$

$$g_{\theta}(v_u; x) = Encoder_{V}(x) \tag{19}$$

$$f_{\theta}(w_i; s) = (1-\lambda) f_{\theta}(w_i; s) + \lambda \cdot g_{\theta}(v_u; x) \tag{20}$$

其中， $f_\theta(w_i;s)$ 表示文本特征向量， $g_\theta(v_u;x)$ 表示视觉特征向量， λ 为门控权重系数[4]。

已有研究表明跨模态检索融合方法在多模态命名实体识别任务上的表现不及其它主流模型，主要原因是目前多模态数据集规模较小，候选检索图像较少而难以充分地与指定文本配对。

综上，尽管多模态融合在多模态命名实体识别研究中取得了不少进展，但仍然面临以下挑战：（1）模态特征对齐问题。目前多数研究主要运用注意力机制实现模态间交互，然而该方式对于多种模态描述同一实体在高层语义是否统一问题上缺乏解释性[20]；（2）模型融合有效性问题。现有研究构建的多模态命名实体识别模型通常将多模态表示、对齐和融合等环节包含在一个深层神经网络中，导致各个环节边界模糊未能充分利用模态间的互补信息[72]；（3）模态噪声问题。多模态融合的数据应该保证语义的强关联性，而在实际任务中难以保证模态间语义的一一对应，难免会引入噪声[72]。

5 多模态实体识别中的预训练模型

预训练模型近些年被广泛应用于命名实体识别任务，通过大规模语料预训练来学习不同模态实体之间的语义对应关系，既提升了视觉语言模型的鲁棒性和泛化性，也节约了大量的计算资源。受此启发，预训练模型相继扩展到多模态场景。相对于单文本的预训练模型，多模态预训练模型可以更好地对细粒度多模态语义单元间相关性进行建模，从而保证模态之间的强关联性[3,11]，这为解决多模态命名实体识别任务中面临的语义鸿沟问题提供了新方案。

多模态预训练一般遵循同一个研究框架，以图像—文本为例，通常包括视觉编码模块、文本编码模块、多模态融合模块、解码模块（可选）和预训练任务五个模块[3,9,11]。图 3 给出了常见的多模态预训练架构流程，首先将各单模态经过相应编码模块得到特征向量表示，接着输入至多模态融合模块进行交互融合，然后根据实际任务需求确定是否需要解码模块，最后执行预训练任务得到最终的多模态特征表示。

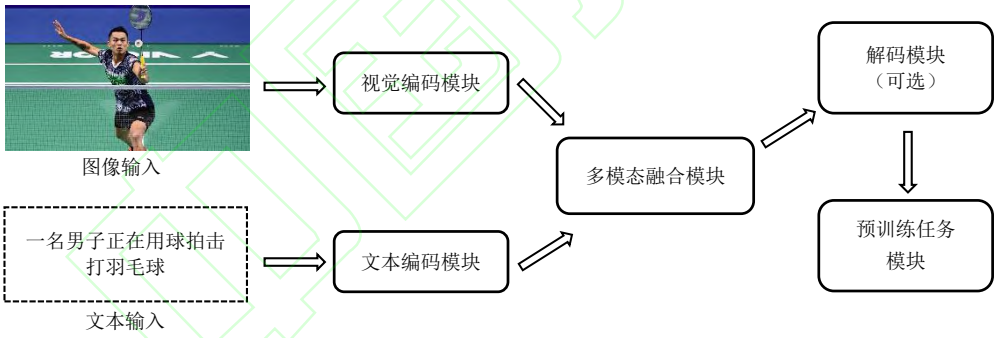


图 3 多模态预训练架构（以图像—文本为例）

Fig 3. Multimodal Pre-Training Architecture (With Image-Text Example)

多模态预训练模型通常分为单流架构和双流架构[11]。单流架构往往仅有单一的编码器，主要应用于检索和理解任务；而双流架构一般由编码器和解码器两部分组成，主要应用于生成和识别任务，两种架构的对比如表 1。

表 1 单流架构和双流架构的比较

Table1 Comparison of Single-Stream And Dual-Stream Architectures

架构种类	原理	优点	缺点	代表性模型
单流架构	将文本和视觉特征组合在一起，馈入单个 Transformer 块，通过合并注意力来融合多模态输入	参数效率更高	无法解耦，须成对送入编码	VisualBERT[73]、VL-BERT[74]、UNITER[75]
双流架构	将文本和视觉特征独立输入到	各模态的网络深度	参数量大	ViLBERT[76]、

	不同的编码块，不共享参数，通常使用交叉注意力用于实现跨模态交互	不同，独立编码，自由组合；可快速解耦	LXMBert[77]、CLIP[78]

目前针对多模态命名实体识别任务的多模态预训练模型相对较少，其中具有代表性的研究有 Sun 等[18,19]提出的 RIVA 和 RpBERT 多模态预训练模型，RIVA 模型通过注意力机制充分利用视觉信息以辅助文本实体识别任务，而 RpBERT 利用关系传播机制解决多模态模型中视觉注意线索存在的噪音问题，两模型在模态输入方式、组成模块和预训练任务方面的对比如表 2。

表 2 多模态命名实体识别中多模态预训练模型
Table2 Multimodal Pre-Training Models in Multimodal Named Entity Recognition

模型	模型输入	组成模块	预训练任务
RIVA[18]	乘法拼接图像和文本模态	1） 图文关系门控网络 2） 注意力引导视觉上下文网络 3） 视觉语言上下文网络	1）图片文本关系分类 2）掩码区域预测
RpBERT[19]	加法拼接图像和文本模态	1） 图文关系分类模块 2） 视觉语言学习模块	1）图片文本关系分类 2）关系传播机制

从表 2 可以看出 RIVA 模型和 RpBERT 模型均使用简单拼接方式融合不同模态数据以作为多模态预训练模型的输入，如式（21）、（22）所示，其中 f_t 和 f_v 分别为文本和图像的特征向量。

$$Input\left(RIVA\right)=f_t\times f_v \tag{21}$$

$$Input\left(RpBERT\right)=Word\ \ Token\ Enbedding+Image\ Block\ \ Embedding \tag{22}$$

RIVA 和 RpBERT 的主要区别是模型的架构设计，RIVA 通过基于门控单元的注意力机制捕捉与文本相关的视觉区域信息以实现图像和文本之间的语义关联；而 RpBERT 增加了关系传播机制实现图像和文本的匹配，通过两个共享参数的多模态 BERT 结构实现文本和图像的深度融合。

尽管多模态预训练模型能够对细粒度语义单元进行建模，但目前多模态预训练模型在命名实体识别上应用较少，主要挑战表现在两个方面：（1）视觉表示方面。现有的视觉语言 BERT 模型如 VisualBERT、VL-BERT 和 UNITER 等[73-75]，均使用目标检测模型生成的感兴趣区域（Region of Interest，RoI）特征表示作为视觉特征表示，RoI 检测目的是降低视觉信息的复杂性，并使用语言线索执行掩蔽区域分类任务[79]。然而对于不相关的文本—图像对，视觉噪音信息可能会增加对语言特征的干扰，Arshad 等[13]通过实验证实了弱相关或不相关的图像会降低模型的预测精度。（2）预训练任务方面。多模态预训练主要包括掩码语言建模（Masked Language Modeling，MLM）、掩码区域预测（Masked Region Prediction，MRP）和图像文本匹配（Image-Text Matching，ITM）三个任务[24,75,79]，MLM 和 MRP 有助于视觉语言模型学习图像—文本之间细粒度的关联特征，而 ITM 则能在细粒度层面上将两种模态的信息进行对齐。然而，目前多模态预训练模型一般是在图像字幕数据集上进行训练，如 COCO 数据集或 Conceptual Captions 数据集[80-82]，这类数据集假设图像和文本是高度匹配的，而该假设在多模态实体识别任务中使用的部分数据集上未必成立。因此，在实际任务中直接将已有模型迁移到多模态命名实体识别任务极有可能导致模型性能的下降[83]，需要动态调整多模态预训练任务。

6 总结与展望

本文对多模态命名实体识别的相关研究工作进行系统梳理，阐述了多模态命名实体识别的相关概念和研究框架，重点对多模态命名实体识别中的多模态特征表示、多模态融合策略和预训练模型进行了分析。针对当前多模态命名实体识别研究存在的不足，提出以下问题和展望。

（1）面向领域的高质量多模态数据集构建问题。由于多模态数据集构建需要投入大量的时间和精力，目前大多数多模态命名实体识别研究主要依赖基于社交平台的图片—文本公开数据集，然而单一的数据来源限制了多模态命名实体识别任务的研究领域，不利于机器对多模态数据的推理和学习，构建面向具体领域的高质量数据集已经成为多模态命名实体识别亟待解决的问题。未来工作在领域文本模态数据基础上引入视频、声音等多模态数据是构建多模态数据集的切实可行方案，进而推动领域多模态命名实体识别研究进展。

（2）多模态融合中异构数据冲突问题。多模态命名实体识别研究在多模态融合阶段存在语义冲突、重复和噪声等问题，解决方案通常是借助注意力机制隐式的对齐多模态语义特征，然而这种方法将多模态语义对齐过程融入一个神经网络，不易对该过程加以主动约束。未来工作一方面可以从细粒度的模态特征提取切入，如通过视觉对象特征提取可有效解决数据稀疏和视觉噪声问题；另一方面，尝试引入多任务学习显式对齐多模态语义特征，如引导模型通过逻辑推理判别模态之间的关联性，进而加深机器对多模态信息的认知、理解和学习能力。

（3）多模态命名实体识别的预训练任务问题。多模态预训练模型使用的多模态数据集中每一组数据组合都是语义高度相关的，而实际应用中难以保证异构数据的语义对应，因此直接将主流的预训练模型迁移至多模态实体识别任务中可能导致性能下降。未来工作可以从预训练任务切入，在通用的多模态预训练模型基础上制定针对命名实体识别的预训练任务，如图文配对等多模态语义对齐任务，目的在于保证模型通过海量数据学习到丰富的多模态语义知识，从而进一步提升多模态命名实体识别效果。

（4）多模态命名实体识别中模态定位问题。已有的多模态命名实体识别研究主要基于文本模态而将其它模态作为文本信息的补充，最终输出文本序列的预测结果。然而，多模态学习的最终目标是使模型具备类似人类多维度获取信息和知识的能力，即各模态间应互相补充语义信息而不是为某一种模态服务。多模态命名实体识别作为多模态学习子任务之一，现阶段工作还未能实现上述目标，因此后续可基于现有工作进一步拓宽多模态命名实体识别研究范畴，构建统一的模型框架实现不同模态实体信息对齐，根据细分领域有针对性的输出具体模态的实体信息。

参考文献：

[1] Moon S, Neves L, Carvalho V. Multimodal Named Entity Recognition for Short Social Media Posts[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018: 852-860.

[2] Zhang Q, Fu J, Liu X, et al. Adaptive Co-Attention Network for Named Entity Recognition in Tweets[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018: 5674-5681.

[3] 吴友政, 李浩然, 姚霆, 等.多模态信息处理前沿综述：应用、融合和预训练[J]. 中文信息学报, 2022, 36(5):1-20.（Wu Youzheng,Li Haoran, Yao Ting, et al. A Survey of Multimodal Information Processing Frontiers: Application, Fusion And Pre-Training[J]. Journal of Chinese Information Processing, 2022, 36(5):1-20）

[4] Yao W, Yoshinaga N. Visually-Guided Named Entity Recognition by Grounding Words with Images via Dense Retrieval[C]//言語処理学会第 28 回年次大会, 2022:1361-1365.

[5] Elliott D, Frank S, Hasler E. Multilingual Image Description with Neural Sequence Models[J]. arXiv preprint arXiv:1510.04709, 2015.

[6] Antol S, Agrawal A, Lu J, et al. VQA: Visual Question Answering[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015:2425-2433.

[7] Zhu X, Li Z, Wang X, et al. Multi-Modal Knowledge Graph Construction and Application: A Survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2015,14(8):1-20.

[8] Baltrušaitis T, Ahuja C, Morency L P. Multimodal Machine Learning: A Survey and Taxonomy[J].

IEEE Transactions on Pattern Analysis and Machine Intelligence,2018, 41(2): 423-443.
[9] Liang P P, Zadeh A, Morency L P. Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions[J]. arXiv preprint arXiv:2209.03430, 2022.
[10] 何俊,张彩庆,李小珍等.面向深度学习的多模态融合技术研究综述[J].计算机工程,2020,46(5):1-11.（He Jun, Zhang Caiqing, LI Xiaozhen, et al. Survey of Research on Multimodal Fusion Technology for Deep Learning[J]. Computer Engineering, 2020, 46(5):1-11.）
[11] 王惠茹,李秀红,李哲等.多模态预训练模型综述[J].计算机应用,2023,43(4):991-1004.（Wang Huiru, Li Xiuhong, Li Zhe, et al. Survey of Multimodal Pre-Training Models[J]. Journal of Computer Applications, 2023,43(4):991-1004.）
[12] Lu D, Neves L, Carvalho V, et al. Visual Attention Model for Name Tagging in Multimodal Social Media[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 1990-1999.
[13] Arshad O, Gallo I, Nawaz S, et al. Aiding Intra-Text Representations with Visual Context for Multimodal Named Entity Recognition[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 337-342.
[14] Yu J, Jiang J, Yang L, et al. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:3342-2252.
[15] Asgari-Chenaghlu M, Feizi-Derakhshi M R, Farzinvash L, et al. CWI: A Multimodal Deep Learning Approach for Named Entity Recognition from Social Media Using Character, Word and Image Features[J]. Neural Computing and Applications, 2022: 1-18.
[16] Zhang D, Wei S, Li S, et al. Multi-Modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(16): 14347-14355.
[17] Zheng C, Wu Z, Wang T, et al. Object-Aware Multimodal Named Entity Recognition in Social Media Posts with Adversarial Learning[J]. IEEE Transactions on Multimedia, 2020, 23: 2520-2532.
[18] Sun L, Wang J, Su Y, et al. RIVA: A Pre-Trained Tweet Multimodal Model Based on Text-Image Relation for Multimodal NER[C]//Proceedings of the 28th International Conference on Computational Linguistics. 2020: 1852-1862.
[19] Sun L, Wang J, Zhang K, et al. Rpbert: A Text-Image Relation Propagation-Based BERT Model for Multimodal NER[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13860-13868.
[20] Liu L, Wang M, Zhang M, et al. UAMNer: Uncertainty-Aware Multimodal Named Entity Recognition in Social Media Posts[J]. Applied Intelligence, 2022, 52(4): 4109-4125.
[21] Sui D, Tian Z, Chen Y, et al. A Large-Scale Chinese Multimodal NER Dataset with Speech Clues[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 2807-2818.
[22] 范涛, 王昊, 陈玥彤. 基于深度迁移学习的地方志多模态命名实体识别研究[J]. 情报学报, 2022, 41(4):412-423.（Fan Tao, Wang Hao, Chen Yuetong. Research on Multimodal Named Entity Recognition of Local History Based on Deep Transfer Learning[J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(4):412-423.）
[23] Xuan Z, Bao R, Jiang S. FGN: Fusion Glyph Network For Chinese Named Entity Recognition[C]//China Conference on Knowledge Graph and Semantic Computing. Springer, Singapore, 2020: 28-40.
[24] Chen D, Li Z, Gu B, et al. Multimodal Named Entity Recognition with Image Attributes And Image Knowledge[C]//International Conference on Database Systems for Advanced Applications.

Springer, Cham, 2021: 186-201.
[25] Wang X, Tian J, Gui M, et al. PromptMNER: Prompt-Based Entity-Related Visual Clue Extraction and Integration for Multimodal Named Entity Recognition[C]//International Conference on Database Systems for Advanced Applications. Springer, Cham, 2022: 297-305.
[26] Xu B, Huang S, Sha C, et al. MAF: A General Matching and Alignment Framework for Multimodal Named Entity Recognition[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022: 1215-1223.
[27] Liu Y, Huang S, Li R, et al. USAF: Multimodal Chinese Named Entity Recognition Using Synthesized Acoustic Features[J]. Information Processing & Management, 2023, 60(3): 103290.
[28] Tian Y, Sun X, Yu H, et al. Hierarchical Self-Adaptation Network for Multimodal Named Entity Recognition in Social Media[J]. Neurocomputing, 2021, 439: 12-21.
[29] Wang X, Gui M, Jiang Y, et al. ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition[J]. arXiv preprint arXiv:2112.06482, 2021.
[30] 李晓腾,张盼盼,勾智楠等.基于多任务学习的多模态命名实体识别方法[J].计算机工程,2023,49(4):114-119.（Li Xiaoteng, Zhang Panpan, Gou Zhinan, et al. Multi-Modal Named Entity Recognition Method Based on Multi-Task Learning[J]. Computer Engineering, 2023, 49(4): 114-119.）
[31] Huang Y, Du C, Xue Z, et al. What Makes Multi-Modal Learning Better than Single[J]. Advances in Neural Information Processing Systems, 2021, 34: 10944-10956.
[32] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv:1301.3781, 2013.
[33] 李代祎,张笑文,严丽.一种基于异构图网络的多模态实体识别方法[J/OL].小型微型计算机系统:1-10[2023-07-24].http://kns.cnki.net/kcms/detail/21.1106.TP.20230711.1048.002.html（Li Daiyi, Zhang Xiaowen, Yan Li. A Multimodal Name Entity Recognition Method Based on Heterogeneous Graph Network[J]. Journal of Chinese Computer Systems: 1-10[2023-07-24].http://kns.cnki.net/kcms/detail/21.1106.TP.20230711.1048.002.html）
[34] Kattenborn T, Leitloff J, Schiefer F, et al. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 173: 24-49.
[35] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
[36] Devlin J, Chang M W, Lee K, et al. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:810.04805,2018.
[37] Chen S, Aguilar G, Neves L, et al. Can Images Help Recognize Entities? A Study of The Role of Images for Multimodal NER[J]. arXiv preprint arXiv:2010.12712, 2020.
[38] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[39] Zhu T, Wang Y, Li H, et al. Multimodal Joint Attribute Prediction and Value Extraction for E-Commerce Product[J]. arXiv preprint arXiv:2009.07162, 2020.
[40] Hu X. Multimodal Named Entity Recognition and Relation Extraction with Retrieval-Augmented Strategy[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 3488-3488.
[41] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),2017,39 (6) :1137–1149.
[42] Kiela D, Bottou L. Learning Image Embeddings Using Convolutional Neural Networks for Improved Multi-Modal Semantics[C]//Proceedings of the 2014 Conference on Empirical Methods

in Natural Language Processing (EMNLP). 2014: 36-45.
[43] Anderson P, He X, Buehler C, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.
[44] Dong L, Xu S, Xu B. Speech-Transformer: A No-Recurrence Sequence-To-Sequence Model for Speech Recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5884-5888.
[45] Purwins H, Li B, Virtanen T, et al. Deep Learning for Audio Signal Processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(2): 206-219.
[46] 胡峰松, 张璇. 基于梅尔频率倒谱系数与翻转梅尔频率倒谱系数的说话人识别方法[J]. 计算机应用,2012, 32(9):2542-2544.（Hu Fengsong, Zhang Xuan. Speaker Recognition Method Based on Mel Frequency Cepstrum Coefficient and Inverted Mel Frequency Cepstrum Coefficient[J]. Journal of Computer Applications, 2012, 32(9):2542-2544.）
[47] Zhang X, Yuan J, Li L, et al. Reducing the Bias of Visual Objects in Multimodal Named Entity Recognition[C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023: 958-966.
[48] Liu P, Li H, Ren Y, et al. A Novel Framework for Multimodal Named Entity Recognition with Multi-level Alignments[J]. arXiv preprint arXiv:2305.08372, 2023.
[49] Khare Y, Bagal V, Mathew M, et al. MMBERT: Multimodal BERT Pretraining for Improved Medical VQA[C]//2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE, 2021: 1033-1036.
[50] Jiang Y G, Wu Z, Wang J, et al. Exploiting Feature and Class Relationships in Video Categorization With Regularized Deep Neural Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(2): 352-364.
[51] Habibian A, Mensink T, Snoek C G M. Video2vec Embeddings Recognize Events When Examples are Scarce[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(10): 2089-2103.
[52] Fukui A, Park D H, Yang D, et al. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding[J]. arXiv preprint arXiv:1606.01847, 2016.
[53] Lu J, Yang J, Batra D, et al. Hierarchical Question-Image Co-Attention for Visual Question Answering[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016:289-297
[54] Zadeh A, Liang P P, Poria S, et al. Multi-Attention Recurrent Network for Human Communication Comprehension[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1): 5642-5649.
[55] Pang L, Ngo C W. Mutlimodal Learning with Deep Boltzmann Machine for Emotion Prediction in User Generated Videos[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 2015: 619-622.
[56] Martínez H P, Yannakakis G N. Deep Multimodal Fusion: Combining Discrete Events and Continuous Signals[C]//Proceedings of the 16th International Conference on Multimodal Interaction. 2014: 34-41.
[57] Rasiwasia N, Costa Pereira J, Coviello E, et al. A New Approach to Cross-Modal Multimedia Retrieval[C]//Proceedings of the 18th ACM International Conference on Multimedia. 2010: 251-260.
[58] Wang B, Yang Y, Xu X, et al. Adversarial Cross-Modal Retrieval[C]//Proceedings of the 25th ACM International Conference on Multimedia. 2017: 154-162.
[59] Wang X, Ye J, Li Z, et al. CAT-MNER: Multimodal Named Entity Recognition with Knowledge-Refined Cross-Modal Attention[C]//2022 IEEE International Conference on Multimedia and Expo

(ICME). IEEE, 2022: 1-6.
[60] Yin Y, Meng F, Su J, et al. A Novel Graph-Based Multi-Modal Fusion Encoder for Neural Machine Translation[J]. arXiv preprint arXiv:2007.08742, 2020.
[61] Poria S, Chaturvedi I, Cambria E, et al. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016: 439-448.
[62] Zadeh A, Zellers R, Pincus E, et al. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.
[63] Zhou B, Zhang Y, Song K, et al. A Span-based Multimodal Variational Autoencoder for Semi-supervised Multimodal Named Entity Recognition[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 6293-6302.
[64] Nojavanasghari B, Gopinath D, Koushik J, et al. Deep Multimodal Fusion for Persuasiveness Prediction[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 284-288.
[65] Kampman O, Barezi E J, Bertero D, et al. Investigating Audio, Visual, and Text Fusion Methods for End-To-End Automatic Personality Prediction[J]. arXiv preprint arXiv:1805.00705, 2018.
[66] Wu Z, Zheng C, Cai Y, et al. Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1038-1046.
[67] Zhao X, Tang B. Multimodal Named Entity Recognition via Co-Attention-Based Method with Dynamic Visual Concept Expansion[C]//International Conference on Neural Information Processing. Springer, Cham, 2021: 476-487.
[68] Kim J H, Jun J, Zhang B T. Bilinear Attention Networks[J]. Advances in neural information processing systems, 2018, 31: 1564-1574.
[69] Zhao G, Dong G, Shi Y, et al. Entity-level Interaction via Heterogeneous Graph for Multimodal Named Entity Recognition[C]//Findings of the Association for Computational Linguistics: EMNLP 2022. 2022: 6345-6350.
[70] 尹奇跃, 黄岩, 张俊格,等. 基于深度学习的跨模态检索[J]. 中国图象图形学报,2021, 26(6):1368-1388.（Yin Qiyue, Huang Yan, Zhang Junge, et al. Survey On Deep Learning Based Cross-Modal Retrieval[J]. Journal of Image and Graphics, 2021, 26(6):1368-1388.）
[71] 李志义, 黄子风, 许晓绵. 基于表示学习的跨模态检索模型与特征抽取研究综述[J]. 情报学报,2018, 37(4):422-435.（Li Zhiyi, Huang Zifeng, Xu Xiaomian. A Review of the Cross-Modal Retrieval Model and Feature Extraction Based on Representation Learning[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(4): 422-435）
[72] 唐樾, 马静. 基于增强对抗网络和多模态融合的谣言检测方法[J]. 情报科学,2022, 40(6):108-114+131.（Tang Yue, Ma Jing. A Rumor Detection Method Based on Enhance Adversarial Network and Multimodal Fusion[J]. Information Science, 2022, 40(6):108-114.）
[73] Li L H, Yatskar M, Yin D, et al. Visualbert: A Simple and Performant Baseline for Vision and Language[J]. arXiv preprint arXiv:1908.03557, 2019.
[74] Su W, Zhu X, Cao Y, et al. Vl-Bert: Pre-Training of Generic Visual-Linguistic Representations[J]. arXiv preprint arXiv:1908.08530, 2019.
[75] Chen Y C, Li L, Yu L, et al. Uniter: Universal Image-Text Representation Learning[C]//European Conference on Computer Vision. Springer, Cham, 2020: 104-120.
[76] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-And-Language Tasks[J]. Advances in Neural Information Processing Systems, 2019, 32:13–23.
[77] Tan H, Bansal M. Lxmert: Learning Cross-Modality Encoder Representations from Transformers[J].

arXiv preprint arXiv:1908.07490, 2019.
[78] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models from Natural Language Supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.
[79] Li G, Duan N, Fang Y, et al. Unicoder-Vl: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11336-11344.
[80] Chen X, Fang H, Lin T Y, et al. Microsoft Coco Captions: Data Collection and Evaluation Server[J]. arXiv preprint arXiv:1504.00325, 2015.
[81] Dou Z Y, Xu Y, Gan Z, et al. An Empirical Study of Training End-To-End Vision-And-Language Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18166-18176.
[82] Yuan L, Cai Y, Wang J, et al. Joint Multimodal Entity-Relation Extraction Based on Edge-enhanced Graph Alignment Network and Word-pair Relation Tagging[J]. arXiv preprint arXiv:2211.15028, 2022.
[83] Wang P, Chen X, Shang Z, et al. Multimodal Named Entity Recognition with Bottleneck Fusion and Contrastive Learning[J]. IEICE Transactions on Information and Systems, 2023, 106(4): 545-555.

通讯作者（Corresponding author）： 韩普（Han Pu），ORCID：0000-0001-5867-4292，E-mail: hanpu@njupt.edu.cn。

基金项目： 本文受国家社科基金项目“面向多模态医疗健康数据的知识组织模式研究”（项目编号：22BTQ096）、江苏高校青蓝工程、南京邮电大学 1311 人才计划和江苏省研究生科研创新计划基金项目“面向多模态医疗健康数据知识图谱构建研究”（项目编号：KYCX23_0930）资助。

This work was supported by the National Social Science Foundation of China under the project of "Research on Knowledge Organization Mode for Multimodal Healthcare Data" (Project No. 22BTQ096), the Blue Project of Jiangsu Province, the 1311 Talent Program of Nanjing University of Posts and Telecommunications, and the Jiangsu Postgraduate Research and Innovation Program under the project of "Study of Constructing a Knowledge Graph for Multimodal Healthcare Data" (Project No. KYCX23_0930).

作者贡献说明：

韩普：提出研究思路，对研究方法提供指导，撰写论文，修改论文；
陈文祺：文献收集整理，撰写论文，修改论文。

利益冲突声明：

所有作者声明不存在利益冲突关系。