计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(2): 294−304, 2026

基于强化策略反馈的多模态自适应实体识别方法

焦明海　樊本航　王　静　彭玉怀
（东北大学计算机科学与工程学院　沈阳　110167）
（mjiao@mail.neu.edu.cn）

A Multimodal Adaptive Entity Recognition Method Based on Reinforcement Strategy Feedback

Jiao MingHai, Fan Benhang, Wang Jing, and Peng Yuhuai
（*School of Computer Science and Engineering*, *Northeastern University*, *Shenyang* 110167）

Abstract　The core objective of the named entity recognition (NER) is to identify entities with specific semantic categories and types in unstructured text. With the rapid growth of social media, textual information is often accompanied by visual content, forming multimodal data. To improve the accuracy of entity recognition, multimodal NER (MNER) methods exploit semantic information from different modalities fully to achieve the complementary and deep fusion of cross-modal features. However, differences in the representation of modalities may introduce visual noise that interferes with entity recognition. Issues such as entity ambiguity or contextual semantic vagueness within the textual modality complicate recognition. To address these challenges, we propose a multimodal NER method based on a reinforcement strategy feedback with adaptive loss mechanism. First, the method adopts a three-stage chain of thought (COT) reasoning process based on GPT-4o to form a progressive reasoning framework that incorporates the adaptive feedback mechanism in reinforcement learning. The degree of matching between images and text is scored, and the interference of visual noise is effectively filtered using an adaptive decision function. Second, four task-specific loss functions are designed and jointly optimized through an adaptive weighted fusion strategy to alleviate the uncertainty caused by contextual ambiguity. Experiments on two representative public datasets (Twitter-2015 and Twitter-2017) show that the overall $F1$ scores of our proposed method are 86.45% and 93.80%, respectively, representing a significant improvement on current state-of-the-art baseline models.

Key words　named entity recognition; unstructured text; multimodal data; reinforcement learning; adaptive loss; chain of thought

摘　要　命名实体识别（named entity recognition, NER）的核心目标是从非结构化文本中识别出具有特定语义类别的实体与类型。随着社交媒体的迅速发展，文本信息往往与视觉信息共同出现，形成多模态内容。为了提升实体识别的准确性，多模态命名实体识别（multi-modal NER, MNER）方法利用不同模态中的语义信息，实现信息互补与深度融合。然而，不同模态之间的表征差异可能引入视觉噪声，干扰实体识别。文本模态中存在实体指代不清或上下文语义模糊的问题，增加了识别难度。针对上述问题，提出了一种基于强化策略反馈与自适应损失机制的MNER方法。首先，该方法采用基于GPT-4o的3阶段思维链（chain of thought, COT）推理流程，形成渐进式推理框架，融合强化学习中的自适应反馈机制，对图像

收稿日期：2025-04-21；**修回日期**：2025-08-07
基金项目：中央高校基础科研基金（N25GFZ015）；中国航空科学基金项目（2020Z066050001）；国家自然科学基金项目（72573032）
This work was supported by the Fundamental Research Funds for the Central Universities (N25GFZ015), the Aeronautical Science Foundation of China (2020Z066050001), and the National Natural Science Foundation of China (72573032).
通信作者：彭玉怀（pengyuhuai@mail.neu.edu.cn）

与文本之间的匹配程度进行评分，并利用自适应决策函数有效过滤视觉噪声的干扰。其次，设计了4类面向具体任务的损失函数，并利用自适应加权融合策略进行优化，以缓解上下文模糊带来的识别不确定性。在2个公开数据集Twitter-2015和Twitter-2017上开展实验，结果表明所提方法的总体$F1$分数分别达到86.45%和93.80%，显著优于当前主流基线模型。

关键词 命名实体识别；非结构化文本；多模态数据；强化学习；自适应损失；思维链

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550273 **CSTR:** 32373.14.issn1000-1239.202550273

命名实体识别(named entity recognition, NER)[1-2]作为自然语言处理(natural language processing, NLP)领域的基石，对于信息检索、智能问答系统以及知识图谱的构建等多方面的应用具有至关重要的意义[3]。随着多模态数据日益丰富，多模态命名实体识别(multimodal named entity recognition, MNER)[4]应运而生，通过高效融合文本与视觉模态数据等跨模态信息，不仅能够提升对非结构化文本数据中实体定位与分类的准确性，而且在文本上下文信息不足或存在语义模糊性的情况下，借助视觉模态提供的额外线索，极大地增强了实体识别的精准度[5]。尤其是在一些特定情境下，如图1文本中的“鸟巢”既可以指代鸟类用于栖息繁殖的巢穴(被视为杂项(miscellaneous, MISC)实体)，也可能是指代著名的体育场馆——国家体育场(视为组织(organization, ORG)实体)。在这种情况下，仅依靠文本信息难以准确区分其实体类型，而结合相关的视觉信息则能有效解决这一问题，实现更加精确的实体识别。然而，由于跨模态异构数据融合过程中可能引入大量的噪声干扰，以及视觉特征中的冗余信息会对文本序列中的非实体部分造成误判，因此精准的MNER方法是确保自然语言处理系统有效性和可靠性不可或缺的一环。



文本: The Bird’s Nest is designed with a modular structure that can withstand strong winds and rainstorms.

Fig. 1 Image information assisted recognition

图1 图像信息辅助识别

在早期阶段，一些研究者通常采用全局图像编码策略来引导文本中单词学习整幅图像的视觉感知表征，然而这种粗粒度对齐方式一方面会引发视觉模态噪声干扰，另一方面会造成细粒度实体特征丢失。随后，越来越多的研究开始关注文本与图像之间的细粒度语义对齐。许多研究者通过优化跨模态注意力模型，在图文特征融合方面取得了不错的效果，推动了跨模态特征之间的信息交互，实现了视觉模态特征在文本实体语义表征中的深度融合。

然而，现有的方法还存在2个问题：1)多模态间的表征差异易引发视觉噪声干扰，当图文语义存在显著偏差时，会严重误导实体识别的准确性。2)文本模态内部存在实体指代上下文模糊，具体表现为实体多义性和实体边界划分的模糊性。

针对上述MNER任务所面临的问题和挑战，我们提出了一种基于强化策略反馈与自适应损失(reinforcement strategy feedback and adaptive loss, RFAL)的MNER方法。具体而言，我们首先使用基于大语言模型(large language model, LLM)的思维链(chain of thought, COT)[6]并结合强化学习(reinforcement learning, RL)[7]中的策略优化思想，通过渐进式推理和自适应反馈精准过滤视觉信息所带来的噪声干扰，并设计基于多类动态损失的自适应机制以解决不同样本歧义特性问题，从而提高实体分类的准确性。

综上所述，本文的贡献有如下3点：

1)提出了一种基于强化策略反馈与自适应损失的MNER分类方法，旨在减少视觉噪声干扰和实体上下文模糊性的影响，在2个层面，如图像与文本、文本内部实现更一致的多模态对齐。

2)分别提出了图文噪声过滤层(image and text noise filtering layer, INF)和自适应多类型损失优化层(adaptive multi-class loss optimization layer, AMLO)这2层协同工作，从而提高实体的分类性能。

3)在2个目前广泛使用的MNER数据集上进行了大量的实验，以证明提出方法的有效性，实验结果表明，本文方法在这2个数据集上达到了SOTA的性能。

1 相关工作

作为跨模态信息处理领域的重要研究方向，MNER旨在通过融合视觉信息与文本信息实现对实体信息的联合抽取，扩展了传统的基于文本的NER任务。相较于传统单模态实体识别方法，MNER需要实现对文本、图像等模态数据的对齐与整合，通过构建跨模态语义表征实现多源信息的语义统一，并基于联合特征对实体解码。如何通过跨模态表征学习实现视觉线索与实体语义的精准对齐，是当前需要面临的巨大挑战。

早期的方法主要通过全局视觉特征编码实现对图文模态的隐式交互。Moon等人[8]设计了基于LSTM-CNN的文本-图像联合建模框架。其采用长短期记忆(LSTM)网络提取文本语义特征，同时运用卷积神经网络(CNN)解析视觉信息，并通过模态注意力机制实现2种模态的信息整合；Zhang等人[9]提出了自适应协同注意力网络，动态调整这2种模态的融合权重。与上述使用全局图像处理方法不同，后续研究则更关注细粒度图文关联的优化方向。这类方法强调提取与文本语义高度相关的视觉区域特征。Lu等人[10]首先提出从图像中筛选出与文本最相关的视觉区域，构建基于注意力模型的跨模态隐式交互机制。Yu等人[11]提出的UMT模型引入了多模态交互模块，通过使用BERT和ResNet预训练模型，分别提取文本与视觉特征，并利用Transformer中Encoder结构实现跨模态语义的深层关联，有效捕捉到图像语义驱动的文本特征与文本引导的视觉特征之间的双向映射关系。Zhang等人[12]构建统一的多模态图结构，通过捕捉文本词元与图像区域节点间的跨模态关联，实现细粒度语义单元之间的交互。

尽管上述研究取得了显著的成果，然而在跨模态数据中融合不同模态的信息导致的模态表征对齐偏差问题尚未得到有效的解决。Xu等人[13]基于对比学习[14]构建了多模态匹配对齐框架，其可以自适应确定视觉信息的保留比例，建立跨模态一致性表征空间。Chen等人[15]提出基于动态门控网络的多模态学习方法，通过构建可学习的阈值单元对视觉特征进行选择性强化，在保留有效视觉特征的同时抑制噪声干扰，并与多模态关系抽取任务进行联合训练，从而得到具有跨模态对齐能力和优异泛化特性的多模态特征。Sun等人[16]将多模态关系传播引入到BERT预训练方法中，通过构建多任务学习框架，在大规模图文数据集上训练和验证关系传播对MNER任务的作用效果，并判别视觉信息与文本内容之间的语义相关性。Zhou等人[17]提出一种基于变分自编码器的多模态特征融合方法，通过构建统一的潜在表征空间实现图文数据的跨模态对齐。

综上所述，上述提到的大部分方法都采用全局图像信息与文本信息相融合，然而这些方法在进行MNER任务时依然存在以下的问题：1)视觉信息和文本信息不匹配会误导模型进行实体的识别，引发视觉信息的噪声干扰。2)文本模态中普遍存在实体上下文模糊，具体表现为多义词歧义和边界模糊的问题，影响实体的识别。针对以上问题，本文提出了RFAL方法，通过对视觉噪声过滤以及自适应多类型损失模块的融合，提升了在MNER任务上的识别性能。

2 RFAL方法

本文所提出的RFAL方法聚焦于跨模态异构数据的细粒度精准匹配，结合跨模态特征的增强交互，实现复杂信息的高效对齐。RFAL方法主要由图文噪声过滤层以及自适应多类型损失优化层2部分组成。本文提出的整体方法框架如图2所示。

MNER旨在通过联合分析文本-图像对中的跨模态语义信息，对文本序列中的实体以及实体类型进行识别与正确分类。给定输入文本序列$S=\{w_1,w_2,\cdots,w_n\}$及其关联图像I，其中w_i表示S中第i个字符，并且$Y=\{y_1,y_2,\cdots,y_n\}$表示文本序列中的所有字符对应的实体类型标签。标签集Y基于BIO2标签模式构建[18]，覆盖4类预定义实体类型：人物(person, PER)、位置(location, LOC)、组织(ORG)和杂项(MISC)，具体包含“O”(非实体)、“B-PER”(人物实体开始)、“I-ORG”(组织实体内部)等细粒度标签。通过将MNER建模为序列标注任务，模型需要对每个实体分配唯一标签，以同时捕获文本语义与视觉模态的潜在关系，最终实现跨模态实体边界识别与类型判断。

2.1 图文噪声过滤层

图文噪声过滤层是基于GPT-4o构建具有3阶段结构的思维链推理流程，形成基于LLM思维链的多阶段递进式推理框架。同时，融合强化学习中的自适应反馈机制，动态评估图像与文本之间的语义匹配程度，借助自适应决策函数，实现噪声数据的多层次精准过滤。通过多阶段的验证体系，每个阶段均

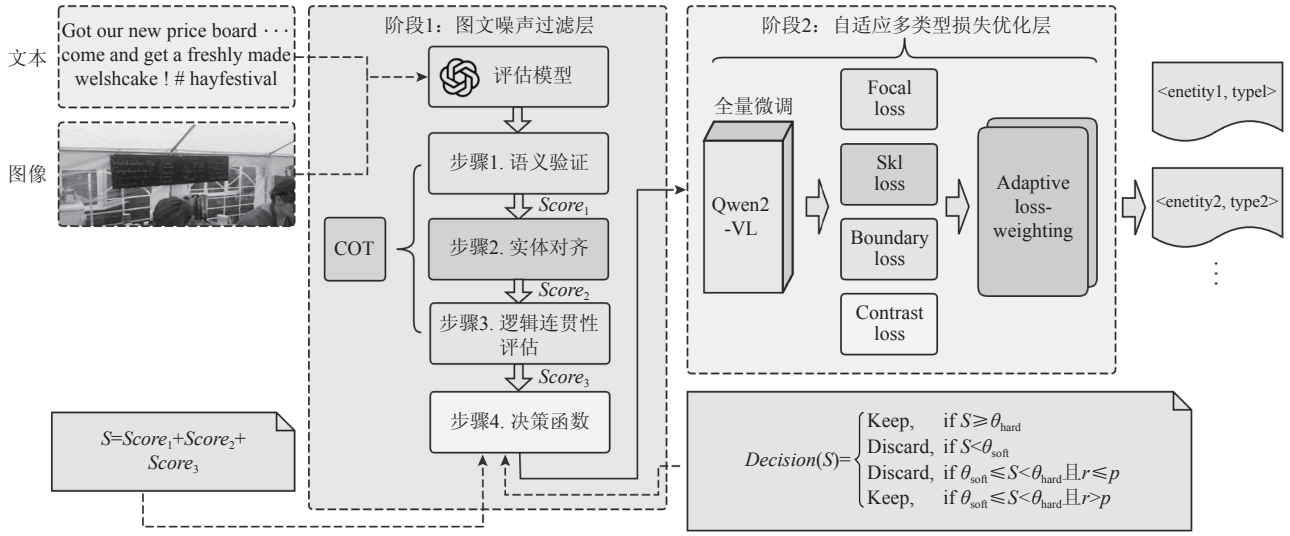


Fig. 2 The structure of RFAL

图 2 RFAL 结构

设置可量化的评价指标与动态调整的阈值参数，解决了 MNER 数据中存在的图文不匹配而引发的噪声干扰问题。

本文提出的基于 COT 的多阶段递进式推理模块，包含语义验证、实体对齐和逻辑连贯评估 3 个评分机制，逐步消除图文噪声对 MNER 任务的影响，并结合动态阈值优化实现自适应决策。

步骤 1. 在语义验证阶段，旨在解决文本描述和图像表征在空间关系上可能存在的显示冲突。由于图文模态在空间方位描述（如“左侧”“上方”）上的显性矛盾是构成初级噪声的主要来源，本步骤通过结构化提示工程[19]引导 GPT-4o 生成具有可解释性的推理过程。具体而言，首先让 GPT-4o 从文本模态中提取空间关系描述，并识别图像中对应物体的位置关系，然后根据图像和文本中空间关系的差异程度输出相应评分。该步骤的输出评分范围设定在 0~40 之间。评分函数为：

$$Score_1=\begin{cases}40, & \text{if } Match(T_{\text{spatial}},I_{\text{spatial}}),\\ 20, & \text{if } PartialMatch(T_{\text{spatial}},I_{\text{spatial}}),\\ 0, & \text{其他},\end{cases} \tag{1}$$

其中 T_{spatial} 为文本模态中的空间关系，I_{spatial} 为图像中的空间关系。

步骤 2. 在实体对齐阶段，针对实体虚构（如文本实体在视觉模态中缺失）及表征模糊是次级噪声的核心来源，传统的二值化判断（存在/不存在）无法有效量化实体存在性概率分布，且忽视不同实体类型对语义完整性的差异贡献。本步骤通过统计显著性加权与置信度评估实现精准化过滤。首先使用词频-逆文档频率（TF-IDF）统计量筛选出关键实体即通过 IDF 降低高频普通实体的权重，突出低频关键实体的权重。TF-IDF 的计算公式为

$$TF\text{-}IDF(e)=\left(\frac{f_e}{N}\right)\times\ln\left(\frac{D}{d_e+1}\right), \tag{2}$$

其中 f_e 为实体频次，N 为文本总词数，D 为总数据量，d_e 为包含实体 e 的数据量。在此基础上，为 4 类实体类型分别赋予不同的权重优先级，例如赋予人物、地点等核心实体类型更高的权重，最后通过归一化加权融合得到最终实体权重。实体权重的函数为

$$w_e=\frac{TF\text{-}IDF(e)\times\lambda_{\text{type}}(e)}{\sum_{e'\in\mathcal{E}}TF\text{-}IDF(e')\times\lambda_{\text{type}}(e')}, \tag{3}$$

其中 w_e 是实体 e 的归一化权重，$TF\text{-}IDF(e)$ 表示实体 e 的 TF-IDF 值，$\mathcal{E}$ 为文本中所有实体的集合，$\lambda_{\text{type}}(e)$ 为实体 e 的类型权重，$\sum_{e'\in\mathcal{E}}TF\text{-}IDF(e')\times\lambda_{\text{type}}(e')$ 为文本中所有实体的加权和。在计算出每个实体权重后，通过结构化提示词（prompt）引导 GPT-4o 生成细粒度存在性置信度评估。模型输出连续概率值 $p_e\in[0,1]$，其中 p_e 量化文本实体在视觉模态中的语义对齐程度（如部分遮挡实体可赋值为 p_e=0.6）。最后进行加权求和得到该步骤得分，该步骤的输出评分范围在 0~40 之间。最终评分计算函数为：

$$Score_2=40\times\sum_{e\in\mathcal{E}}(w_e\times p_e), \tag{4}$$

其中 w_e 为每个实体的权重，p_e 为 GPT-4o 对图像中是否存在实体 e 的置信度。

步骤 3. 在逻辑连贯性评估中，主要是解决文本描述与视觉表征在事件发展逻辑及因果关联层面的

隐性矛盾问题。尽管前序阶段已确保实体存在性与空间关系的局部一致性,但仍需要确保跨模态语义的全局连贯性以及实体之间的现实逻辑性。因此,本步骤提出基于双重因果图匹配的评估框架,核心在于构建文本与图像双模态的因果图结构,并通过多维相似度度量实现语义连贯性量化。具体而言,首先通过结构化 prompt 引导 GPT-4o 分别从文本和图像中提取关键实体和实体状态来构成因果图中的节点集合: $V_{\text{text|image}}=\{v_i|v_i=(e_i,s_i)\}$, 其中 e_i 表示实体, s_i 表示实体状态。节点相似度的计算如式(5)所示。

$$Sim_{\text{node}}=\frac{\left|V_{\text{text}}\cap V_{\text{image}}\right|}{\left|V_{\text{text}}\cup V_{\text{image}}\right|}, \tag{5}$$

其中 V_{text} 表示从文本中提取的节点集合, V_{image} 表示从图像中提取的节点集合。在边关系建模层面,分别构建文本和图像的边集合: $E_{\text{text|image}}=\{(v_p,r_q,v_r)\}$, 其中 r_q 表示因果谓词(如“导致”“影响”等)。边的相似度计算由边的结构相似度与边的语义相似度共同决定。边的结构相似度使用 Jaccard 通过比较边拓扑连接模式计算,边的语义相似度则由 GPT-4o 根据文本和图像的边集合关系输出[0,1]之间的相似概率,边的相似度计算公式如式(6)所示。

$$\begin{aligned}Sim_{\text{edge}}=&\alpha\times Jaccard\left(E_{\text{text}},E_{\text{image}}\right)+\\&(1-\alpha)\times GPT_{\text{sim}}\left(E_{\text{text}},E_{\text{image}}\right),\end{aligned} \tag{6}$$

其中 α 为人工预先设置的权重超参数, $Jaccard(E_{\text{text}},E_{\text{image}})$ 为边的结构相似度, $GPT_{\text{sim}}(E_{\text{text}},E_{\text{image}})$ 为边的语义相似度。

最终该步骤的评分通过融合节点与边相似度的几何平均合成,输出评分范围在 0~20 之间。最终评分计算函数为

$$Score_3=\sqrt{Sim_{\text{node}}\times Sim_{\text{edge}}}\times 20。 \tag{7}$$

步骤 4. 在完成多阶段评分计算后,最终会通过层次化决策函数实现图文匹配度的最终判定。该决策机制使用动态边界阈值与概率化决策相结合的混合范式。具体而言,首先随机抽取 100 条由人工审核的黄金标注样本构建基准评分分布: 均值(u)、标准差(σ)。然后计算总评分 $S=Score_1+Score_2+Score_3$, 以此建立自适应阈值计算模型。阈值动态调整过程遵循渐进式策略优化原则,定义硬决策边界和软决策边界。硬决策阈值: $\theta_{\text{hard}}=\mu-k\sigma$, 其中严格系数 $k\in[0.5,1.5]$, 该阈值确保显著噪声样本的确定性过滤。软决策阈值: $\theta_{\text{soft}}=\theta_{\text{hard}}-\Delta$, 其中缓冲区间 $\Delta=0.2\sigma$, 用于缓解决策函数在临界区域的敏感性。总体的决策函数设计为:

$$Decision(S)=\begin{cases}\text{Keep, if } S\geqslant\theta_{\text{hard}},\\ \text{Discard, if } S<\theta_{\text{soft}},\\ \text{Discard, if } \theta_{\text{soft}}\leqslant S<\theta_{\text{hard}} \text{ 且 } r\leqslant p,\\ \text{Keep, if } \theta_{\text{soft}}\leqslant S<\theta_{\text{hard}} \text{ 且 } r>p,\end{cases} \tag{8}$$

其中 $r\sim U(0,1)$ 为均匀分布随机变量, $p=1-\exp(-\beta(S-\theta_{\text{soft}}))$ 为自适应图像丢弃概率,衰减系数 β 控制决策边界的平滑度。通过在决策过程中引入概率化扰动,防止模型陷入局部最优,并且指数衰减函数确保接近硬阈值的样本具有更高的保留概率。

2.2 自适应多类型损失优化层

在社交媒体场景中的 MNER 任务中,实体的上下文模糊主要表现为 2 种典型特征: 多义词歧义(如“Java”指编程语言或岛屿)和边界模糊性(如“New York”被截断为“York”)。针对传统固定权重的损失函数难以动态适应不同样本歧义特性的局限性而导致模型对复杂语境泛化能力不足的问题,本节提出一种基于动态损失的自适应机制,通过对 Qwen2-VL-7B 多模态大模型(multimodal large language model, MLLM)进行全量微调[20],并融合多维度精细化设计的损失函数,从而实现具有上下文感知能力的自适应优化机制。

基于对目标数据集的分析,我们发现该数据集中包含 4 类具有显著差异的实体类型。因此,我们提出假设: 针对不同的实体类型设计差异化的损失函数可能会优化模型的分类效果,为了验证该假设,本文分别构建了 4 类针对性的损失函数。并且我们注意到不同的损失函数对各类实体的适用性可能存在差异,为避免单一损失函数主导优化过程,我们引入了自适应权重优化机制,在训练过程中,权重参数 α 会进行动态调整。通过观察测试集上的消融实验结果,验证了本节提出的方法的有效性。

社交媒体文本中的实体类型分布呈现显著的长尾特性。这种分布不均衡导致传统的交叉熵损失在训练过程中易被高频类别主导以及低频类别实体表征学习不充分。此外,部分实体因上下文模糊成为“困难样本”,其预测概率接近决策边界($p=0.5$),此类样本在标准交叉熵损失中的梯度贡献较低、优化效率不足。基于上述分析我们设计了类别平衡的焦点交叉熵损失函数。焦点交叉熵损失通过双重调节机制增强对困难样本与低频类别的关注。

$$L_{\text{focal}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\alpha_c(1-p_{i,c})^{\gamma}y_{i,c}\ln p_{i,c}, \tag{9}$$

其中 N 为批次实体数量, C 为实体类别数, $y_{i,c}$ 为真实

标签，$p_{i,c}$为模型输出的预测概率，α_c为类别权重（根据类别频率动态调整），γ为调节因子。相比于传统交叉熵，该损失在训练后期持续优化困难样本，提升决策边界附近的分类精度。

为缓解模型对错误标签的过度自信而导致的泛化性不足问题，本文融合了对称标签平滑的 KL 损失函数。损失函数为

$$L_{\text{skl}}=\frac{1}{2N}\sum_{i=1}^{N}\left[D_{\text{KL}}(\boldsymbol{q}_i\|\boldsymbol{p}_i)+D_{\text{KL}}(\boldsymbol{p}_i\|\boldsymbol{q}_i)\right], \tag{10}$$

其中$\boldsymbol{q}_i$为平滑后的标签分布，$q_{i,c}=(1-\epsilon)y_{i,c}+\frac{\epsilon}{C}$，$\epsilon$为超参数，通常取值为$\epsilon=0.1$，$C$为实体类别数。$\boldsymbol{p}_i$为模型的预测概率分布。$D_{\text{KL}}$为 KL 散度，$D_{\text{KL}}(\boldsymbol{q}_i\|\boldsymbol{p}_i)=\sum_{i\in C}\boldsymbol{q}_i\ln\frac{\boldsymbol{q}_i}{\boldsymbol{p}_i}$。通过进一步分析发现，实体边界的模糊性通常会导致 2 种典型的错误：不完整检测（如将“New York City”截断为“York”）与嵌套冲突（如“Bank of America”包含子实体“America”），传统的方法在预测时未充分考虑边界位置的合理性，可能直接导致类型误分类。为此，我们进一步融合了隐式边界一致性损失，其利用实体类型标签隐式学习实体边界信息。该损失函数设计为

$$L_{\text{boundary}}=\frac{1}{T}\sum_{t=1}^{T}(\lambda_1 d_t\times I(y_t\neq y_{t+1})-\lambda_2 d_t\times I(y_t=y_{t+1})), \tag{11}$$

其中T为文本的序列长度，λ_1和λ_2为平衡权重，$d_t=\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_1$表示相邻 token 的 L1 距离，$\boldsymbol{p}_t\in\mathbb{R}^C$表示第 t 个 token 的预测概率分布，$I(\cdot)$为指示函数（条件满足为 1，否则为 0）。该损失函数在计算时，若相邻 token 类型不同，则增大差异，否则减小差异。

最后，考虑到之前设计的损失函数缺少对同类实体的全局语义约束，我们通过对比学习的思想使同类实体的预测分布相近，异类实体的预测分布远离，从而增强模型对上下文语义的敏感度。该损失函数的计算公式为

$$L_{\text{contrast}}=-\frac{1}{N}\sum_{i=1}^{N}\ln\frac{\exp\left(\frac{s(\boldsymbol{p}_i,\boldsymbol{m}_{c_i})}{\tau}\right)}{\sum_{c\in C}\exp\left(\frac{s(\boldsymbol{p}_i,\boldsymbol{m}_c)}{\tau}\right)}, \tag{12}$$

其中N为批次内实体数量，$\boldsymbol{p}_i\in\mathbb{R}^C$表示第$i$个实体的预测概率分布，$\boldsymbol{m}_{c_i}$为类型$c$的类中心，$\boldsymbol{m}_c=\frac{1}{|B_c|}\sum_{j\in B_c}\boldsymbol{p}_j$，$s(\boldsymbol{p}_i,\boldsymbol{m}_{c_i})$表示 2 个向量间的余弦相似度，$\tau$为温度系数。

在完成 4 类损失函数的设计后，本文使用一种基于自适应权重优化[21]的动态融合策略。通过在模型微调过程中自适应调整其权重数值分布，实现对 4 类损失函数的平衡优化，具体设计如下：

$$L_{\text{total}}=\alpha_1 L_{\text{focal}}+\alpha_2 L_{\text{skl}}+\alpha_3 L_{\text{boundary}}+\alpha_4 L_{\text{contrast}}, \tag{13}$$

其中$\alpha_1,\alpha_2,\alpha_3,\alpha_4$为自适应优化的权重参数，采用高斯分布进行初始化，初始值为$\alpha_i\sim\mathcal{N}(0,0.1^2)$，在模型微调过程中，权重参数通过反向传播自适应优化，并且其满足非负性与归一化约束：$\sum_{i=1}^{4}\alpha_i\geqslant 0$，$\alpha_i=1$。

3 实验与分析

3.1 数据集与评估指标

为验证本文所提出方法的有效性，使用了从社交媒体上收集的、目前广泛使用的 Twitter-2015[15]，Twitter-2017[9] 两个数据集进行了相关实验，模型只有在正确识别出文本中的实体跨度以及其对应的实体类型才算完全正确的识别，表明了本文方法在 MNER 任务上的有效性。其中 Twitter-2015 数据集包括 4 000/1 000/3 357 的训练/验证/测试分割，Twitter-2017 数据集包括 3 373/723/723 的训练/验证/测试分割。2 个数据集中的每一条数据都可能包含多个实体，并且由文本和其对应的图像组成。其中实体类型主要包含 4 类：人物（PER）、位置（LOC）、组织（ORG）、其他（MISC）。2 种数据集统计数据如表 1 所示。

Table 1 Statistical Data of Two Datasets

表 1 2 个数据集的统计数据

类型	Twitter-2015			Twitter-2017		
	训练集	验证集	测试集	训练集	验证集	测试集
PER	2 217	552	1816	2 943	626	621
LOC	2 091	522	1 697	731	173	178
ORG	928	247	839	1 674	357	395
MISC	940	225	726	701	150	157
总计	6 176	1 546	5 078	6 049	1 324	1 351
推文数量	4 000	1 000	3 257	3 373	723	723

本文通过计算 2 个数据集的 $F1$ 分数、召回率 R（recall）、精确率 P（precision），来评估我们的模型在 MNER 任务中的效果。在当前任务中，精确率和召回率会呈现负相关的关系，理想情况下期望实现两者的同步优化，但是在实际的应用场景中，召回率的提升会使精确率下降。为了有效平衡这 2 个指标，采用 $F1$ 分数作为综合评价指标，该指标更能客观地反映出模型的性能。$F1$ 分数在需要双重验证模型效果的大多数自然语言处理任务中具有重要的评估价

值。相应的计算如式(14)(15)(16)所示。

$$P=\frac{TP}{TP+FP}, \tag{14}$$

$$R=\frac{TP}{TP+FN}, \tag{15}$$

$$F1=2\times\frac{P\times R}{P+R}, \tag{16}$$

其中 TP 为模型正确识别出文本中的实体跨度和其对应实体类型的实体数量, FP 为模型错误识别实体跨度和实体类型的实体数量, FN 为模型未识别出的实体数量。

3.2 超参数设置与实验环境

本文提出的模型通过 PyTorch 框架在 4 张 Nvidia GeForce RTX 4090 GPU 上进行了 MNER 的相关实验, 使用 Qwen2-VL-7B 基座多模态大模型进行全量微调。模型的所有超参数配置如表 2 所示。

Table 2 Training Hyper-Parameters

表 2 训练的超参数

超参数	取值
batch_size	4
epoch	6
learning_rate	1E−5
warmup_ratio	0.03
bf16	True

3.3 对比方法

为了验证本文提出的 RFAL 方法在 MNER 任务上的效果, 本文与在 Twitter-2015 和 Twitter-2017 数据集上所提出的一些经典的 MNER 方法进行了对比, 包含以下所提出的方法:

1)BiLSTM-CRF[22]。该方法提出将双向长短时记忆网络(bidirectional long short-term memory, BiLSTM)与条件随机场(conditional random field, CRF)相结合用于序列标注任务。通过 BiLSTM 编码器有效捕获文本的上下文特征, 并利用 CRF 层实现实体序列的全局最优解码。

2)CNN-BiLSTM-CRF[23]。该方法在 BiLSTM-CRF 架构基础上引入 CNN, 通过 CNN 与 BiLSTM 提取文本特征, 并利用 CRF 实现实体序列的全局最优解码, 是一种经典的序列标注方法。

3)BERT-CRF[22]。该方法通过 BERT 捕捉文本深层语义表征, 并采用 CRF 层替代传统 softmax 分类器进行最终的标签预测。

4)MAF[13]。该方法通过动态调节视觉信息保留比例与跨模态噪声抑制机制, 提升了多模态表征的鲁棒性。

5)BGA-MNER[24]。该方法提出双向生成对齐方法, 通过图像到文本和文本到图像的生成任务, 直接学习实体与图像对象间的隐含关系, 无需对应标注, 有效缓解模态偏差问题。

6)PGIM[25]。该方法提出 2 阶段框架, 利用 ChatGPT 作为隐式知识库, 通过多模态相似示例引导生成辅助精炼知识, 增强文本实体预测。该方法在 MNER 任务中表现出色, 具有更强的鲁棒性和泛化能力。

7)DPE-MNER[26]。该方法提出动态融合多模态信息的迭代推理框架, 通过“分解、优先、消除”策略分层融合多粒度文本和图像信息, 逐步优化实体识别结果。

8)MPMRC-MNER[27]。该方法提出基于多模态提示的机器阅读理解框架, 将文本查询转化为包含图像信息的多模态提示, 并通过对比学习优化跨模态语义空间, 提升 MNER 性能。

9)CMAT[28]。该方法提出一种多任务多模态学习架构, 通过跨模态辅助任务助力 MNER。该方法利用跨模态匹配和互信息最大化分离共享特征与任务特异性特征, 并通过自适应融合优化跨模态特征。

3.4 实验结果与分析

本文所提出的 RFAL 方法与上述其他方法在 Twitter-2015 数据集和 Twitter-2017 数据集上的效果对比如表 3 所示。

首先, 通过表 3 中的对比结果可以看出基于预训练语言模型的 BERT-CRF 框架在 2 个测试数据集上的效果相较于传统 BiLSTM-CRF 等方法实现了 $F1$ 分数上的 7 个百分点左右的提升。原因在于基于海量文本预训练的 BERT 模型通过深度上下文建模捕获多维语义信息, 显著提升了语义消歧能力。并且 CRF 层的序列约束机制利用标签转移概率优化, 有效增强相邻标记间的关联一致性。

其次, 通过实验观察发现, 融合视觉模态的 MNER 模型在 2 个测试数据集上相较于单一文本模态的 NER 模型效果上表现出更优的性能。这表明, 在社交媒体常见的短文本语境下, 融合视觉信息更有助于实体及其类型的识别。

最后, 本文提出的 MNER 方法在 2 个测试数据集上相较于其他的方法都取得了显著的性能提升, 并且极大地缓解了因图文不匹配而带来的噪声干扰问题以及实体上下文模糊等问题, 证明了 RFAL 的有效性。与基于文本的单模态命名实体识别方法(BERT-CRF)相比, RFAL 在 Twitter-2015 和 Twitter-

Table 3 Comparison of Experimental Results

表 3 实验结果对比 %

方法	Twitter-2015							Twitter-2017						
	PER	LOC	ORG	MISC	*P*	*R*	*F*1	PER	LOC	ORG	MISC	*P*	*R*	*F*1
BiLSTM-CRF	76.77	72.56	41.33	26.8	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80	78.76	79.37
BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
MAF	84.67	81.18	63.35	41.82	71.86	75.10	73.42	91.51	85.80	85.10	68.79	86.13	86.38	86.25
BGA-MNER	86.8	83.62	63.6	42.65	78.60	74.16	76.31	93.71	85.55	85.71	71.05	87.71	87.71	87.71
PGIM	88.34	84.22	**70.15**	52.34	79.21	79.45	79.33	96.46	**89.89**	89.03	**79.62**	90.86	92.01	91.43
DPE-MNER	87.31	84.36	65.93	48.48	76.86	78.27	77.56	92.37	87.47	86.33	73.83	88.46	87.34	87.9
MPMRC-MNER	85.88	83.06	66.6	42.99	77.15	75.39	76.26	92.80	86.80	83.40	72.79	87.1	87.16	87.13
CMAT	86.11	81.62	63.53	41.18	73.22	75.63	74.39	92.02	86.12	85.25	70.20	87.3	86.89	87.09
RFAL（本文）	**90.45**	**86.91**	70.01	**55.28**	**88.85**	**84.19**	**86.45**	**96.82**	88.27	**89.85**	75.38	**94.72**	**92.89**	**93.80**

注：加黑数值表示在当前指标下的最优值。

2017 数据集上的总体 *F*1 分数分别提高了 14.64 个百分点和 10.36 个百分点，并且在各个实体类型上的 *F*1 分数也都远远高于 BERT-CRF。证明了 RFAL 能够有效融合与文本模态相关的视觉模态信息。除此之外，在与最近提出的多模态 SOTA 方法 PGIM 和 DPE-MNER 的比较中，RFAL 在 2 个数据集上的效果要好于它们。具体而言，与 PGIM 相比，在 Twitter-2015 数据集和 Twitter-2017 数据集上的总体 *F*1 分数分别提高了 7.12 个百分点和 2.37 个百分点。与 DPE-MNER 相比，2 个数据集上的总体 *F*1 分数分别提高了 8.89 个百分点和 5.9 个百分点。并且如图 3 所示，在 Twitter-2015 数据集上的效果提升远好于 Twitter-2017 数据集上的效果，这也说明当前的方法受不同数据分布的影响，泛化性不足。而 RFAL 方法在 2 个数据集上的表现都比较优异，这也证明了 RFAL 方法中的 2 个优势：1)图文噪声过滤层有效减少了因图文不匹配而带来的额外视觉信息噪声干扰问题，2)自适应多类型损失优化层有效解决了因文本中实体语义多歧而带来的语义模糊问题。

3.5 消融实验

为了验证本文所提出的 RFAL 方法中的 2 层，即图文噪声过滤层和自适应多类型损失优化层的有效性，进行了如下的消融实验。具体分析如下：

如表 4 所示，其中 w/o I&A 表示消除图文噪声过滤层和自适应多类型损失优化层，w/o INF 表示消除图文噪声过滤层，w/o AMLO 表示消除自适应多类型损失优化层。可以观察到 RFAL 方法在 2 个数据集上的 *F*1 分数上明显优于其他方法，证明了 RFAL 的所有模块对提升 MNER 任务的性能都很重要。

与 RFAL 相比，消除图文噪声过滤层导致 Twitter-

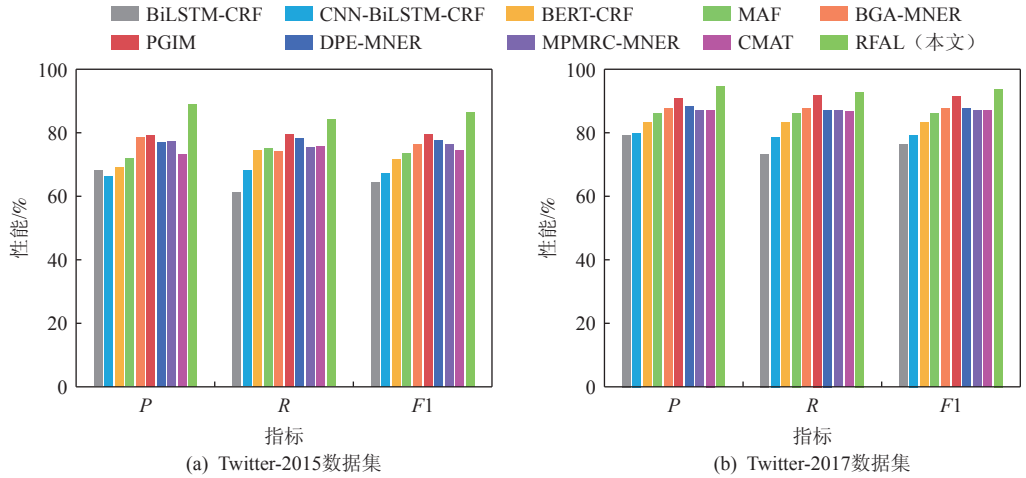


Fig. 3 Comparison of experimental results on two datasets

图 3 2 个数据集上的实验结果对比

Table 4 Ablation Experiments of Model Structure

表 4 模型结构消融实验 %

方法	Twitter-2015			Twitter-2017		
	P	*R*	*F*1	*P*	*R*	*F*1
w/o I&A	88.63	79.65	83.9	93.84	87.93	90.79
w/o INF	88.73	82.43	85.47	95.18	92.08	93.6
w/o AMLO	89.99	80.15	84.78	95.67	90.01	92.75
RFAL（本文）	88.85	84.19	86.45	94.72	92.89	93.8

2015 和 Twitter-2017 在 *F*1 分数上分别下降了 0.98 个百分点和 0.2 个百分点，结果证明了引入图文噪声过滤层可以很好地缓解图文不匹配带来的噪声干扰问题。

此外，消除自适应多类型损失优化层导致 2 个数据集上的 *F*1 分数下降了 1.67 个百分点和 1.05 个百分点，证明了本文在第 2 节提出的假设的有效性可以很好缓解实体上下文模糊的问题。

在消除图文噪声过滤层和自适应多类型损失优化层后，可以观察到同时消除这 2 层会导致最终结果显著下降。这进一步证明了本文所提出的两大方法层对 MNER 任务是有效且十分重要的。

为了进一步验证本文所提出方法中自适应多类型损失优化层中各个组件的有效性，进行了如下的消融实验。

如表 5 所示，可以观察到在图文噪声过滤层的基础上分别加入各类损失函数之后，模型在 2 个数据集上的 *F*1 分数上均有提升，这说明不同类型的损失函数可以有效解决 NER 中普遍存在的实体类型呈现长尾特性、模型对错误标签过渡自信而导致泛化性不足、实体边界模糊以及上下文语义不敏感等问题，验证了本文所提出的自适应多类型损失优化层中 4 类具体任务损失函数的有效性，其能够针对当前 NER 任务中存在的问题发挥作用。

Table 5 Ablation Experiments of Different Components in an Adaptive Multi-Type Loss Optimization Module

表 5 自适应多类型损失优化模块中各组件消融实验 %

方法	Twitter-2015			Twitter-2017		
	P	*R*	*F*1	*P*	*R*	*F*1
INF	89.99	80.15	84.78	95.67	90.01	92.75
INF+L1	89.18	81.74	85.30	95.05	90.90	92.92
INF+L1+L2	89.22	82.59	85.78	95.23	91.71	93.44
INF+L1+L2+L3	89.21	83.28	86.14	94.93	91.49	93.18
RFAL（本文）	88.85	84.19	86.45	94.72	92.89	93.8

注：INF 表示只使用图文噪声过滤层，INF+L1，L2，L3 分别表示在图文噪声过滤层的基础上加入自适应多类型损失优化层中的焦点交叉熵损失函数、标签平滑的 KL 损失函数、隐式边界一致性损失函数。RFAL 是在 INF+L1+L2+L3 的基础上加入类似对比学习的损失函数。

3.6 案例分析

本文从基准测试集中选取了 2 个具有代表性的测试样本，用来比较不同方法的实体识别效果。

当文本模态中存在信息偏差或缺失时，融合视觉特征往往能展现出更好的效果。这是因为图像信息可以补充文本描述的盲区，从而为实体识别任务提供有价值的线索。如表 6(a)中，在纯文本模态下，BERT-CRF 无法正确识别出“皇马”是一个组织类型。然而在结合视觉信息后，多模态模型能正确识别出该实体，并判断出其实体类型。

Table 6 Entity Prediction

表 6 实体预测

方法	视觉信息重要性		图像噪声干扰问题
			
	(a) [Sergio Ramos][1]: [Real Madrid][2] as motivated as two years ago		(b) My [Future Hearts][1] Tour concert was one year ago today
BERT-CRF	1-PER√	2-MISC×	1-MISC√
MAF	1-PER√	2-ORG√	1-PER×
RFAL（本文）	1-PER√	2-ORG√	1-MISC√

注：“√”表示该类型识别正确，“×”表示该类型识别错误。

图像噪声干扰问题存在的主要原因包括：1)非描述性视觉因素的存在(如抽象插画或隐喻图像)；2)跨模态差异化(即视觉符号与文字描述呈现事件的维度不同)。在此类情景下，MNER 方法可能要比单模态方法(如 BERT-CRF)效果更差。在表 6(b)中提供的图像会干扰模型对实体的识别，而多模态方法 MAF 并没有对图像噪声问题进行过滤，因此会将文本中“Future Hearts”的类型“MISC”错误识别为“PER”。RFAL 方法由于充分考虑到图像不匹配而引发的噪声问题，以及实体上下文模糊的问题，所以正确识别了上述的所有实体以及其实体类型。

4 结 论

针对目前 MNER 任务中存在的由于多模态间的表征差异而引发的视觉噪声干扰问题以及文本内部存在的实体指代上下文模糊的问题，本文提出了基于强化策略反馈与自适应损失的 MNER 方法。具体地，本文首先使用基于 GPT-4o 的 COT 方式，并融合

强化学习的自适应反馈机制，实现对视觉噪声的精准过滤；然后设计了动态自适应加权损失层，根据实体上下文模糊问题，针对性地采取多类损失函数优化模型实体识别的性能，提升了最终的实验结果。在2个公开基础的数据集上的实验结果表明，本文提出的MNER的方法明显优于目前大多数方法。

在未来的工作中，计划推进以下两大方面：1）针对跨模态交互中的特征匹配问题，拟构建一种基于多层次特征对齐机制的深度神经网络架构。2）将目前的方法拓展至其他多模态任务，包括但不限于多模态情感分析、多模态人机对话等。

作者贡献声明：焦明海指导论文撰写，提供技术支持；樊本航提出研究思路，设计和实施实验，撰写论文；王静完成了相关文献梳理、实验数据整理，并完善实验方案；彭玉怀完成了论文框架设计、整体内容规划。

参 考 文 献

[1] Mai Chengcheng, Liu Jian, Qiu Mengchuan, et al. Pronounce differently, mean differently: A multi-tagging-scheme learning method for Chinese NER integrated with lexicon and phonetic features[J]. Information Processing & Management, 2022, 59(5): 103041

[2] He Qizhen, Wu Liang, Yin Yida, et al. Knowledge-graph augmented word representations for named entity recognition[C]//Proc of the 34th Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 7919−7926

[3] Wang Jiacheng, Wang Kai, Wang Haofen, et al. Noise detection for distant supervised named entity recongnition[J]. Journal of Computer Research and Development, 2024, 61(4): 916−928 (in Chinese)
(王嘉诚, 王凯, 王昊奋, 等. 面向远程监督命名实体识别等噪声检测[J]. 计算机研究与发展, 2024, 61(4): 916−928)

[4] Arshad O, Gallo I, Nawaz S, et al. Aiding intra-text representations with visual context for multimodal named entity recognition[C]//Proc of the IEEE Int Conf on Document Analysis and Recognition (ICDAR). Piscataway, NJ: IEEE, 2019: 337−342

[5] Zhang Xin, Yuan Jingling, Li Lin, et al. Reducing the bias of visual objects in multimodal named entity recognition[C]//Proc of the 16th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2023: 958−966

[6] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824−24837

[7] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning[J]. arXiv preprint, arXiv: 1312.5602, 2013

[8] Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts[J]. arXiv preprint, arXiv: 1802.07862, 2018

[9] Zhang Qi, Fu Jinlan, Liu Xiaoyu, et al. Adaptive co-attention network for named entity recognition in tweets[C]//Proc of the 32nd Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 5674−5681

[10] Lu D, Neves L, Carvalho V, et al. Visual attention model for name tagging in multimodal social media[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 1990−1999

[11] Yu Jianfei, Jiang Jing, Yang Li, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 3342−3352

[12] Zhang Dong, Wei Suzhong, Li Shoushan, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//Proc of the 35th Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 14347−14355

[13] Xu Bo, Huang Shizhou, Sha Chaofeng, et al. MAF: A general matching and alignment framework for multimodal named entity recognition[C]//Proc of the 15th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2022: 1215−1223

[14] Bhattacharjee A, Karami M, Liu H. Text transformations in contrastive self-supervised learning: A review[J]. arXiv preprint, arXiv: 2203.12000, 2022

[15] Chen Xiang, Zhang Ningyu, Li Lei, et al. Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction[J]. arXiv preprint, arXiv: 2205.03521, 2022

[16] Sun Lin, Wang Jiquan, Zhang Kai, et al. RpBERT: A text-image relation propagation-based BERT model for multimodal NER[C]//Proc of the 35th Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 13860−13868

[17] Zhou Baohang, Zhang Ying, Song Kehui, et al. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition[C]//Proc of the 2022 Conf on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: ACL, 2022: 6293−6302

[18] Sang E F, Veenstra J. Representing text chunks[C]//Proc of the 9th Conf on European Chapter of the Association for Computational Linguistics (EACL). Stroudsburg, PA: ACL, 1999: 173−179

[19] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877−1901

[20] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the Annual Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Strouds-

burg, PA: ACL, 2019: 4171−4186

[21] Yao Ruilin, Xiong Shengwu, Zhao Yichen, et al. Visualgrounding with multi-model conditional adaptation[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 3877−3886

[22] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint, arXiv: 1508.01991, 2015

[23] Ma Xuezhe, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[J]. arXiv preprint, arXiv: 1603.01354, 2016

[24] Chen Feng, Liu Jiajia, Ji Kaixiang, et al. Learning implicit entity-object relations by bidirectional generative alignment for multimodal NER[C]//Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 4555−4563

[25] Li Jinyuan, Li Han, Pan Zhuo, et al. Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge[J]. arXiv preprint, arXiv: 2305.12212, 2023

[26] Zheng Zihao, Zhang Zihan, Wang Zexin, et al. Decompose, prioritize, and eliminate: Dynamically integrating diverse representations for multimodal named entity recognition[C]//Proc of the 2024 Joint Int Conf on Computational Linguistics, Language Resources and Evaluation (LREC-COLING' 2024). Stroudsburg, PA: ACL, 2024: 4498−4508

[27] Bao Xigang, Tian Mengyuan, Zha Zhiyuan, et al. MPMRC-MNER: A unified MRC framework for multimodal named entity recognition based multimodal prompt[C]//Proc of the 32nd ACM Int Conf on Information and Knowledge Management. New York: ACM, 2023: 47−56

[28] Chen Zhengjie, Zhang Yu, Mi S. Assisting multimodal named entity recognition by cross-modal auxiliary tasks[J]. Pattern Recognition Letters, 2023, 175: 52−58



Jiao MingHai, born in 1973. PhD, associate professor, master supervisor. Member of CCF. His main research interests include artificial intelligence, image recognition, and big data parsing and optimization.

焦明海, 1973 年生。博士，副教授，硕士生导师。CCF 会员。主要研究方向为人工智能、图像识别、大数据分析与优化。



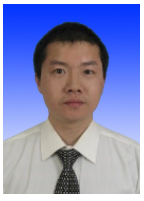
Fan Benhang, born in 2002. Master candidate. His main research interests include machine learning and multimodal information extraction.

樊本航, 2002 年生。硕士研究生。主要研究方向为机器学习、多模态信息抽取。



Wang Jing, born in 1992. PhD candidate. Her main research interests include Internet of things and digital twin.

王　静, 1992 年生。博士研究生。主要研究方向为物联网、数字孪生。



Peng Yuhuai, born in 1983. PhD, professor, PhD supervisor. His main research interests include Internet of things, cyber physical systems, and intelligent information networks.

彭玉怀, 1983 年生。博士，教授，博士生导师。主要研究方向为物联网、网络物理系统、智能信息网络。