◆人工智能算法与应用◆

多模态语义协同交互的图文联合命名实体识别方法*

钟维幸，王海荣**，王　栋，车　淼
（北方民族大学计算机科学与工程学院，宁夏银川　750021）

摘要：针对现有多模态命名实体识别（Multimodal Named Entity Recognition，MNER）研究中存在的噪声影响和图文语义融合不足问题，本文提出一个多模态语义协同交互的图文联合命名实体识别（Image-Text Joint Named Entity Recognition，ITJNER）模型。ITJNER 模型加入图像描述作为额外特征丰富了多模态特征表示，图像描述可以帮助过滤掉从图像特征中引入的噪声并以文本形式总结图像语义信息；还构建了多模态协同交互的多模态语义融合模型，可以加强多模态信息融合，并减少图像信息的语义偏差。在 Twitter-2015 和 Twitter-2017 数据集上进行方法实验，分析实验结果并与 AdaCAN、UMT、UMGF、Object-AGBAN 等方法进行对比。相较于对比方法中的最优方法 UMGF，本方法在 Twitter-2017 数据集上的准确率、召回率、F1 值分别提高了 0.67%、2.26%、0.93%；在 Twitter-2015 数据集上，召回率提高了 0.19%。实验结果验证了本方法的有效性。

关键字：多模态命名实体识别　图文数据　多模态注意力　图像描述　语义融合

中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1005-9164(2022)04-0681-10
DOI：10.13656/j.cnki.gxkx.20220919.008

自媒体的广泛应用致使互联网上的海量数据呈现图像、文本、视频等多模态交融态势，这些数据具有语义互补性，因此，多模态数据的知识抽取和应用成为研究热点，作为基础任务的多模态命名实体识别（Multimodal Named Entity Recognition，MNER）方法研究受到关注。

MNER 领域的初期工作旨在将图像信息利用起来以提升命名识别的效果，通过将单词与图像区域对齐的方式，获取与文本相关的有效视觉上下文。Esteves 等[1]首次在 MNER 任务中使用了视觉信息，将图文联合命名实体识别带入研究者的视野。随后，Zhang 等[2]提出了一种基于双向长短时记忆（Long Short-Term Memory，LSTM）网络模型（BiLSTM）和共注意力机制的自适应共注意网络，这是首个在 MNER 研究上有突出表现的工作。同年，Moon 等[3]、Lu 等[4]也相继提出自己的 MNER 方法，前者

收稿日期：2022-03-24
* 宁夏自然科学基金项目（2020AAC03218），北方民族大学校级科研项目（2021XYZJK06）和北方民族大学研究生创新项目（YCX21092）资助。
【作者简介】
钟维幸（1998－），男，在读硕士研究生，主要从事知识图谱与大数据知识抽取研究。
【**通信作者】
王海荣（1977－），女，博士，副教授，主要从事大数据知识工程研究，E-mail：bmdwhr@163.com。
【引用本文】
钟维幸，王海荣，王栋，等. 多模态语义协同交互的图文联合命名实体识别方法[J]. 广西科学，2022，29(4)：681-690.
ZHONG W X，WANG H R，WANG D，et al. Image-Text Joint Named Entity Recognition Method Based on Multi-modal Semantic Interaction [J]. Guangxi Sciences，2022，29(4)：681-690.

提出一个通用的注意力模块用于自适应地降低或增强单词嵌入、字符嵌入和视觉特征权重,后者则提出一个视觉注意模型,以寻找与文本内容相关的图像区域。在之前的工作中仅用单个单词来捕捉视觉注意,该方式对视觉特征的利用存在不足,Arshad 等[5]将自注意力机制扩展到捕获两个词和图像区域之间的关系,并引入门控融合模块,从文本和视觉特征中动态选择信息。但是在 MNER 中融合文本信息和图像信息时,图像并不总是有益的,如在 Arshad 等[5]和 Lu 等[4]的工作中均提及不相关图像所带来的噪声问题,因此,如何在 MNER 中减小无关图像的干扰成为研究重点。如 Asgari-Chenaghlu 等[6]扩展设计了一个多模态 BERT 来学习图像和文本之间的关系。Sun 等[7,8]提出一种用于预测图文相关性的文本图像关系传播模型,其可以帮助消除模态噪声的影响。为了缓解视觉偏差的问题,Yu 等[9]在其模型中加入实体跨度检测模块来指导最终的预测。而 Liu 等[10]则结合贝叶斯神经网络设计一种不确定性感知的 MNER 框架,减少无关图像对实体识别的影响。Tian 等[11]提出分层自适应网络(Hierarchical Self-adaptation Network,HSN)来迭代地捕获不同表示的子空间中更多的跨模态语义交互。

上述方法学习了粗粒度的视觉对象与文本实体之间的关系。但粗粒度特征可能会忽略细粒度视觉对象与文本实体之间的映射关系,进而导致不同类型实体的错误检测。为此,一些研究开始探索细粒度的视觉对象与文本实体之间的关系。Zheng 等[12]提出一种对抗性门控双线性注意神经网络,将文本和图像的不同表示映射为共享表示。Wu 等[13]提出一种针对细粒度交互的密集协同注意机制,它将对象级图像信息和字符级文本信息相结合来预测实体。Zhang 等[14]提出一种多模态图融合方法,充分利用了不同模态语义单元之间的细粒度语义。除了直接利用图像的原始信息,一些额外信息的加入也有益于 MNER 任务,如 Chen 等[15]在其模型中引入图像属性和图像知识,Chen 等[16]则将图像的描述作为丰富 MNER 的上下文的一种方法。

当前,MNER 仍面临两大挑战:一是无关的图像信息带来的噪声干扰,二是图文语义交互中有效语义信息的丢失。为此,本文提出一种新的多模态语义协同交互的图文联合命名实体识别(Image-Text Joint Named Entity Recognition,ITJNER)模型,引入图像描述以增强视觉数据的特征表示,建立多注意力机制耦合的多模态协同交互模块,通过多个跨模态注意力机制实现模态间语义的充分交互并过滤错误图像所带来的噪声信息,实现图文联合下命名实体的有效识别。

1 方法模型

ITJNER 模型通过协同表示学习图像、文本的深层特征,使用自注意力、跨模态注意力、门控机制通过协同交互的方式实现跨模态语义交互,并加入条件随机场,利用标签间的依赖关系得到最优的预测标签序列。具体模型如图 1 所示。图 1 展示了本方法的核心处理流程,其主要包含多模态特征表示、多模态协同交互与序列标注两个核心模块。

2 多模态特征表示

对图像与文本进行多模态特征表示是图文联合命名实体识别工作的基础,大量研究表明,将文本表示和视觉表示作为多模态特征相结合,可以提高语义提取任务的性能[17,18]。为方便描述对图文特征的抽取与表示工作,将图文对数据集形式化地表示为

$$D=\{I,S\}_{n=1}^{N}, \tag{1}$$

其中,I 为图像,S 为文本,N 为图像-文本数。

2.1 文本特征抽取与表示

对文本特征的抽取是命名实体识别任务的基本,更加轻量化且不影响性能的模型有助于降低后续从算法模型到应用落地的难度,因此本文采用 ALBERT 模型[19]对文本进行特征提取。ALBERT 是一个轻量级的 BERT 模型,其参数比 BERT-large 更少且效果更好,为了降低参数量和增强语义理解能力,其引入词嵌入矩阵分解和跨层参数共享策略,并使用句子顺序预测(Sentence Order Prediction,SOP)任务替换原先的下一句预测(Next Sentence Prediction,NSP)任务。在模型中使用多层双向 Transformer 编码器对输入序列进行编码,其模型结构见图 2。图 2 展示了 ALBERT 模型的核心结构,包含输入层、编码层、输出层,其中每一个 Trm 对应一个 Transformer 编码器。

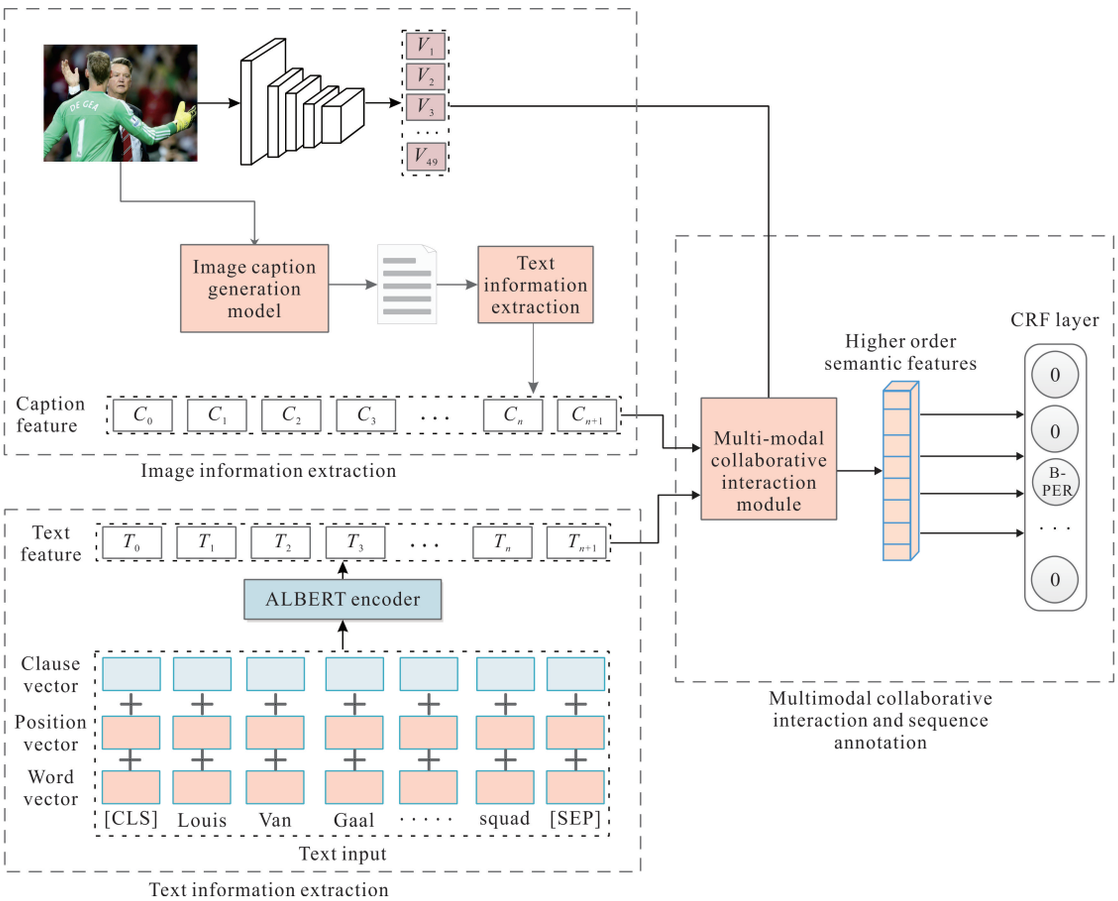

图 1　图文联合命名实体识别模型的整体架构

Fig. 1　Overall architecture of a image-text joint named entity recognition model

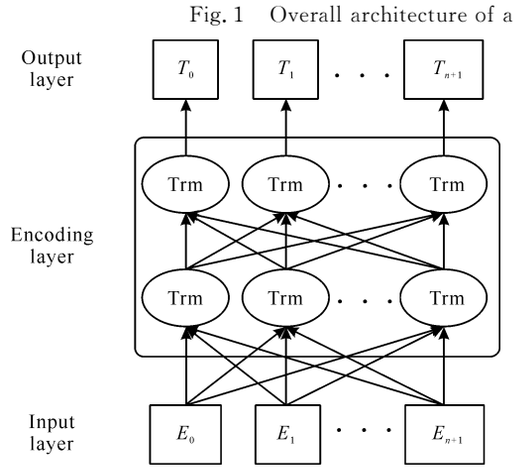

图 2　ALBERT 模型结构图

Fig. 2　Structure diagram of ALBERT model

由于数据集文本可能存在无用的特殊字符，需要对数据进行预处理，对每个输入句子 S 进行标记处理，对不存在的字符使用[UNK]替代，并分别在每个句子的开头和结尾插入两个特殊的标记即[CLS]和[SEP]。形式上，设 $S=[S_0,S_1,S_2,\cdots,S_{n+1}]$ 为修改后的输入句子，其中 S_0 和 S_{n+1} 表示插入的两个令牌。设 $E=[E_0,E_1,E_2,\cdots,E_{n+1}]$ 为句子 S 的标记表示，其中 E_i 为字符向量、分段向量和位置向量的和。将 E 作为 ALBERT 编码层的输入。

$$T=\mathrm{ALBERT}(E), \tag{2}$$

$T=[T_0,T_1,T_2,\cdots,T_{n+1}]$ 为模型的输出向量，其中 $T_i\in\mathbb{R}^d$ 为 E_i 生成的上下文感知表示，d 是向量的维数。在获得文本特征表示的同时，对图像与图像描述特征进行特征抽取。

2.2　图像及图像描述特征的抽取与表示

2.2.1　图像特征抽取

卷积神经网络（Convolutional Neural Networks，CNN）的最新研究进展显示，更强的多尺度表示能力可以在广泛的应用中对图像特征的提取带来性能增益，因此本文采用预训练过的 Res2Net[20] 来提取图像特征。Res2Net 在粒度级别表示多尺度特征，并增加了每个网络层的感受野，相比于传统 ResNet 网络，其在不增加计算复杂度的情况下，提高了网络的特征表示能力。更深层次的网络已经被证明对视觉任务具有更强的表示能力[21]，在综合考虑模型的性能与模型训练效率后，本文最终选择采用 101

层的 Res2Net（Res2Net-101）用于图像特征的提取与表示。

不同图文对数据中的图像大小可能不同,因此首先将它们的大小统一缩放为 224×224 像素,并经随机剪切、归一化等图像预处理方法进行数据增强;然后将调整后的图像输入 Res2Net-101,如式(3)所示。

$$U=\text{Res2Net}(I),I\in D。\tag{3}$$

本文在预训练的 Res2Net-101 中保留了最后一个卷积层输出,以表示每幅图像,遵循大部分研究对卷积核大小的设置,经 Res2Net 进行特征抽取后,获得 7×7=49 个视觉块特征 $U=(u_1,u_2,\cdots,u_{49})$,其中 u_i 是第 i 个视觉块,由 2 048 维向量表示。在将图文特征输入多模态协同交互模块前需保持图文特征向量的维度一致,因此对视觉块特征 U 应用线性变换得到 $V=(v_1,v_2,\cdots,v_{49})$,如式(4)所示。

$$V=W_u^{\mathrm{T}}U,\tag{4}$$

其中,$W_u\in\mathbb{R}^{2048\times d}$ 是一个权重矩阵。

2.2.2　图像描述特征抽取

为了加强图像与文本间的语义融合,本文加入图像描述,并将其视为图文间的过渡信息特征,描述可以帮助过滤掉从图像特征中引入的噪声,同时也可以更好地总结图像的语义。本文使用包含视觉注意力的编解码框架的描述生成模型来生成图像描述,如图 3 所示。

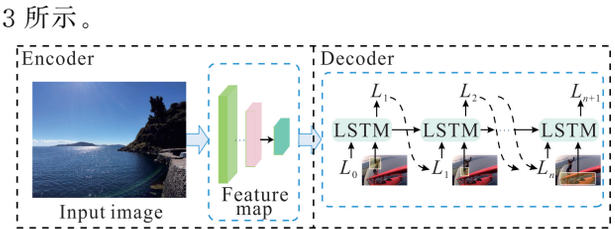

图 3　图像描述模型

Fig.3　Image description model

使用图像特征提取到的视觉块特征 U 作为长短时记忆(LSTM)网络的输入,LSTM 网络通过动态地选择图像特征,提取句子内部单词之间的句法特征、单词位置编码信息,学习图像特征与句法特征、单词特征之间的映射关系,同时加入注意力机制,赋予不同视觉区域以不同的权重,以此缓解视觉噪声干扰。将加权图像特征输入 LSTM,将图像信息逐字转换为自然语言,输出目标为

$$L=[L_0,L_1,L_2,\cdots,L_{n+1}],L_i\in\mathbb{R}^k\tag{5}$$

其中 k 是词汇表的大小,n 是描述句的长度,L_i 代表句子中的第 i 个单词。再将描述 L 作为输入,使用 ALBERT 编码器,得到 $C=[C_0,C_1,C_2,\cdots,C_{n+1}]$,其中 $C_i\in\mathbb{R}^d$ 是 L_i 生成的上下文表示,d 是向量的维数。在得到多模态表示后将其作为协同交互模块的输入,实现多模态特征的语义交互。

3　多模态协同交互与序列标注

多模态协同交互模块获取图像、文本、图像描述特征,利用图像引导进行文本模态融合、文本引导进行图像模态融合,实现不同特征的语义交互,减少视觉偏差。图 4 展示了多模态协同交互模块的具体框架结构,其中包括了以文本向量为键值的跨模态注意力、以图像向量为键值的跨模态注意力、以原始文本向量为键值的非标准自注意力、视觉门控机制。

如图 4 所示,在 ALBERT 模型得到的输出后添加一个标准的自注意力层,以获得每个单词的文本隐藏层表示 $R=(r_0,r_1,\cdots,r_{n+1})$,其中 $r_i\in\mathbb{R}^d$ 为生成的文本隐藏层表示。对图像描述特征 C 和视觉块特征 U 线性变换所得的视觉块特征 V 各添加一个标准自注意力层,分别得到图像描述与图像的隐藏层表示:

$$O=(o_0,o_1,o_2,\cdots,o_{n+1}),\tag{6}$$

$$W=(w_1,w_2,\cdots,w_{49}),\tag{7}$$

其中 $o_i\in\mathbb{R}^d$ 为生成的图像描述隐藏层表示,$w_i\in\mathbb{R}^d$ 为生成的图像隐藏层表示。

3.1　图像引导的文本模态融合

如图 4 左侧所示,为了利用相关图像学习更好的文本表示,本文采用多头跨模态注意力机制,先利用图像描述来引导文本融合,将 $O\in\mathbb{R}^{d\times(n+1)}$ 作为查询,将 $R\in\mathbb{R}^{d\times(n+1)}$ 作为键和值,将 m 设为多头数:

$$A_i(O,R)=\text{softmax}(\frac{[W_{qi}O][W_{ki}R]^{\mathrm{T}}}{\sqrt{d_k}})[W_{ki}R];\tag{8}$$

$$\text{MHA}(O,R)=W_O[A_1(O,R),\cdots,A_m(O,R)]^{\mathrm{T}},\tag{9}$$

其中 A_i 指跨模态注意力的第 i 个头,MHA 表示多头注意力,$\{W_{qi},W_{ki},W_{vi}\}\in\mathbb{R}^{d/m\times d}$ 和 $W_0\in\mathbb{R}^{d\times d}$ 分别表示查询、键、值和多头注意力的权重矩阵。在跨模态注意层的输出后堆叠前馈网络和层归一化等,另外 3 个子层得到描述感知文本表示 $P=(p_0,p_1,\cdots,p_{n+1})$,如式(10)-(11)所示:

$$\tilde{P}=\text{LN}(O+\text{MHA}(O,R)),\tag{10}$$

$$P=\text{LN}(\tilde{P}+\text{FFN}(\tilde{P})),\tag{11}$$

其中FFN表示前馈网络,LN表示层归一化。在利

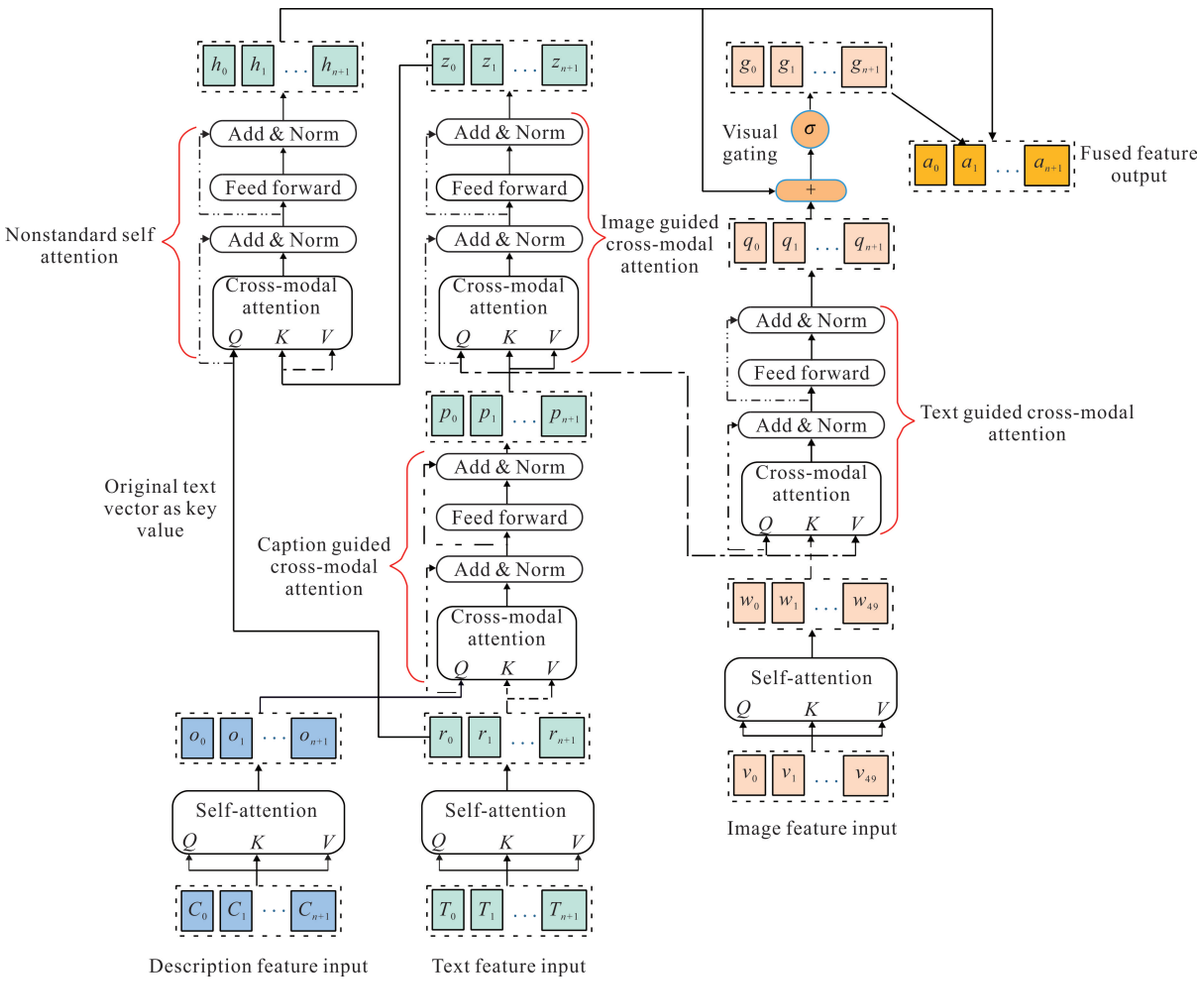

图 4　多模态协同交互模块的框架结构

Fig. 4　Frame structure multimodal cooperative interaction module

用图像描述填补了文本与相关图像间的语义空白后，再利用图像与描述感知文本做跨模态注意力，将 $W \in \mathbb{R}^{d \times 49}$ 作为查询，将 $P \in \mathbb{R}^{d \times (n+1)}$ 作为键和值，与文本和描述的融合方法相似，叠加 3 个子层后输出 $Z = (z_1, z_2, \cdots, z_{49})$，由于以视觉表示作为查询，所以生成的向量 z_i 都对应于第 i 个视觉块，而非第 i 个输入字符，因此另外加入一个跨模态注意力层，以文本表示 R 作为查询，并将 Z 作为键和值，该跨模态注意力层生成最终的图像感知文本表示 $H = (h_0, h_1, \cdots, h_{n+1})$。

3.2　文本引导的图像模态融合

为了将每个单词与其密切相关的视觉块对齐，加入跨模态注意力层为视觉块分配不同的注意力权重。将 P 作为查询，W 作为键和值。与图像引导的文本模态融合对称，文本引导的图像模态融合会生成具有单词感知能力的视觉表示，用 $Q = (q_0, q_1, \cdots, q_{n+1})$ 表示。

相关图像中，部分文本中的一些视觉块可能与单词没有任何关联，同时，文本中的一些单词如助词、数词等也与视觉块少有关联。因此，本文应用一个视觉门控来动态控制每个视觉块特征的贡献，如式(12)所示：

$$g = \sigma((W_h)^{\mathrm{T}} H + (W_q)^{\mathrm{T}} Q), \tag{12}$$

其中 $\{W_h, W_q\} \in \mathbb{R}^{d \times d}$ 是权重矩阵，σ 是元素级的 S 型激活函数。基于动态视觉门控，得到最终的文本感知视觉表示为 $G = (g_0, g_1, \cdots, g_{n+1})$。

在得到最终的图像感知文本表示 H 和最终的文本感知视觉表示 G 后，本文将 H 和 G 拼接，得到图像与文本最终融合的隐藏层表示 $A = (a_0, a_1, \cdots, a_{n+1})$，其中 $a_i \in \mathbb{R}^{2d}$。

3.3　标签依赖的序列标注

在命名实体识别任务中，输出标签对其邻域有着强依赖性，如 I-LOC 不会出现在 B-PER 后。多模态协同交互只考虑了图文对数据中上下文的信息，而没

有考虑标签间的依赖关系,因此,本文添加了一个条件随机场(Conditional Random Field,CRF)来标记全局最优序列,并将隐藏层表示 A 转化为最佳标记序列 $y=(y_0,y_1,\cdots,y_{n+1})$,CRF 可以有效提升此类任务的性能。本文对给定的输入句子 S 及其关联图像 I 的标签序列 y 计算如下:

$$P(y\mid S,I)=\frac{\exp(\text{score}(A,y))}{\sum_{y'}\exp(\text{score}(A,y'))};\quad(13)$$

$$\text{score}(A,y)=\sum_{i=0}^{n}T_{y_i,y_{i+1}}+\sum_{i=1}^{n}E_{a_i,y_i};\quad(14)$$

$$E_{a_i,y_i}=w_{\text{ITJNER}}^{y_i}\cdot a_i,\quad(15)$$

score (A,y)为特征得分,由过渡得分和发射得分两部分组成,其中 $T_{y_i,y_{i+1}}$ 是从标签 y_i 到标签 y_{i+1} 的过渡分数,E_{a_i,y_i} 是标签 y_i 的发射分数;$w_{\text{ITJNER}}^{y_i}\in\mathbb{R}^{2d\times C}$ 是 y_i 特有的权重参数,其中 C 是类数。本文使用最大条件似然如式(16),来学习使对数似然最大化的最佳参数。

$$L(p(y\mid S))=\sum_i\log p(y\mid S)。\quad(16)$$

经上述学习得到全局最优标注序列。

4 验证实验及结果分析

4.1 数据集和方法验证

为验证本文提出的方法,使用 python 语言,利用 pytorch 等技术在 Ubuntu 系统上搭建实验环境,在 Twitter-2015 和 Twitter-2017 两个公共数据集上进行实验,数据集信息如表 1 所示。

对于实验中比较的每种单模态和多模态方法,考虑到文本数据的实际输入范围,将句子输入的最大长度设置为 128。考虑到训练速度的内存大小,将批处理大小设置为 8。对于本方法,对预训练语言模型的参数设置大多数遵循原始论文设置。使用 ALBERT-Base 模型进行文本抽取初始化,使用预训练的 Res2Net-101 来初始化视觉表示,并在训练中保持大小固定。对于多头自注意力层和多头跨模态注意力层,考虑训练效率与精度,在经过调整训练后使用 12 个头和 768 个隐藏单元。同时,经过对超参数多次微调,将学习率、dropout 率和权衡参数 λ 分别设置为 5e-5,0.1 和 0.5,可以在两个数据集的开发集上获得最好的性能。

表 1 数据集详情

Table 1 Dataset details

实体类别 Entity type	Twitter-2015			Twitter-2017		
	Train	Dev.	Test	Train	Dev.	Test
Person	2 217	552	1 816	2 943	626	621
Location	2 091	522	1 697	731	173	178
Organization	928	247	839	1 674	375	395
Miscellaneous	940	225	726	701	150	157
Total	6 176	1 546	5 078	6 049	1 324	1 351
Twitter quantity	4 000	1 000	3 257	3 373	723	723

本实验使用召回率(Recall)、准确率(Precision)、F1 值作为实验评价指标,与 HBiLSTM-CRF-GVATT[5]、BERT-CRF-GVATT[5]、AdaCAN-CNN-BiLSTM-CRF[3] 等 12 种方法的对比结果如表 2 所示。

表 2 对比实验结果

Table 2 Comparison of experimental results

模态 Modal	模型 Model	Twitter-2015			Twitter-2017		
		Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Text	BiLSTM-CRF	68.14	61.09	64.42	79.42	73.43	76.31
	LSTM-CNN-CRF	66.24	68.09	67.15	80.00	78.76	79.37
	HiBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37
	BERT-softmax	68.30	74.61	71.32	82.19	83.72	82.95
	BERT-CRF	71.00	73.27	72.10	82.98	84.46	83.71
	BERT-BiLSTM-CRF	71.03	73.57	72.27	83.20	84.68	83.93
	HBiLSTM-CRF-GVATT	73.96	67.90	70.80	83.41	80.38	81.87
	BERT-CRF-GVATT	69.15	74.46	71.70	83.64	84.38	84.01
	AdaCAN-CNN-BiLSTM-CRF	72.75	68.74	70.69	84.16	80.24	82.15
	AdaCAN-BERT-CRF	69.87	74.59	72.15	85.15	83.20	84.10

续表

Continued table

模态 Modal	模型 Model	Twitter-2015			Twitter-2017		
		Precision（%）	Recall（%）	F1（%）	Precision（%）	Recall（%）	F1（%）
Image + Text	AGBAN	74.13	72.39	73.25	85.36	84.56	85.23
	MDA-CRF	72.81	70.33	71.55	82.64	83.24	83.45
	CWI-Attention	72.37	70.05	71.19	84.56	84.78	83.42
	MT-BERT-CRF	70.48	74.80	72.58	84.60	84.16	84.42
	UMT-BERT-CRF	71.67	75.23	73.41	85.28	85.34	85.31
	MSB-Small-CRF	**74.97**	72.04	73.47	85.20	83.60	84.32
	UMGF	74.49	75.21	**74.85**	86.54	84.50	85.51
	UAMNer	71.78	74.63	73.10	84.13	85.71	84.90
	ITJNER（This study）	73.20	**75.40**	73.61	**87.21**	**86.76**	**86.44**

Note：Bold data in the table indicates the highest score in the current comparison method

4.2 对比实验

实验结果表明，图文联合方法通常可以获得更好的性能，本文方法在 Twitter-2017 数据集上的准确率、召回率、F1 值较对比方法中的最优方法 UMGF 分别提高了 0.67%、2.26% 和 0.93%；在 Twitter-2015 数据集上，召回率提高了 0.19%。

对于单模态方法，预训练的方法明显优于传统的神经网络。例如，BERT-CRF 在 Twitter-2017 数据集上准确率、召回率、F1 值的表现比 HiBiLSTM-CRF 分别高出 0.29%、6.3% 和 3.34%，表明预训练模型在 NER 中具有明显的优势。使用 CRF 解码的 BERT-CRF 的性能优于使用 softmax 的 BERT-softmax，说明 CRF 层对 NER 的有效性。通过对比单模态与多模态方法，可以看到多模态方法的性能明显优于单模态方法。例如，加入视觉门控注意力后，在两个数据集上 HBiLSTM-CRF 较之前的 F1 值分别提高了 1.63% 和 1.5%。此外，相较于 AGBAN、UMT-BERT-CRF 等未使用图像描述的模型，本文方法的性能表现更好，表明结合图像描述有助于完成 NER 任务。

针对本文方法在 Twitter-2015 数据集上表现不佳的情况，本文对数据集的内容进行分析，统计两个数据集的实体分布状态，通过对比图文间实体分布的不同，反映出数据集的图文关联程度，并人工抽样统计数据集的图文关联度，如图 5 所示。

从图 5 可以看到数据集中文本实体分布与图像实体分布之间的差异，图像实体与文本实体并不是完全对应的，图像中的实体对象总量一般会多于其对应的文本所含的命名实体数量，这一差别也体现了数据集中图像文本对之间存在无关联或弱关联情况。对比数据集的图文内容后发现，Twitter-2015 中图文无关联或弱关联现象比 Twitter-2017 中更多，而对本文所提出的方法，图像描述与图像本身有着更高的关联性，因此，在图文无关联或弱关联的图文对数据中，图像描述与文本的语义差距会更大，这也意味着在进行命名实体识别时，带入了无关的噪声数据。由此分析，本文提出的加强图文间融合的方法可以为图文存在相关性的 MNER 带来益处，但对于图文显著无关的情况仍有待改进。

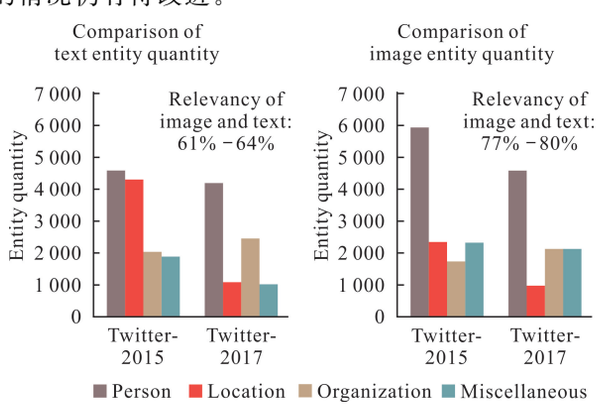

图 5 数据集实体量对比图

Fig. 5 Comparison diagram of dataset entity quantity

4.3 消融实验

为了研究本文图文联合命名实体识别模型中模块的有效性，对模型的核心部件进行消融实验。如表 3 所示，图像描述、视觉门控、图像感知文本融合均对模型生效起重要影响，在去掉图像描述后，模型在 Twitter-2017 数据集上的表现明显变差，而在 Twitter-2015 数据集上的表现却并没有下滑甚至略有提

升,这佐证了 4.2 节的观点,即加入图像描述所带来的影响会因图文数据关联度不同而不同,图文间关联度更大,可以为 NER 任务提供帮助;若图文间关联度不足则可能会起到相反的作用。在多模态协同交互模块中,去除图像感知文本表示后性能明显下降,显示它对模型有不可或缺的作用。而去除视觉门控也会导致轻微的性能下降,这体现了它对整个模型有着一定的重要性。

表 3 消融实验

Table 3 Ablation experiment

模型 Model	Twitter-2015			Twitter-2017		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
ITJNER (This study)	**73.20**	75.30	73.61	**87.21**	**86.76**	**86.44**
w/o image caption	72.33	**75.60**	**73.93**	85.30	85.05	85.17
w/o visual gate	72.53	74.92	73.70	85.51	85.95	85.73
w/o image-aware text-fusion	70.75	74.42	72.54	83.57	84.23	83.89

Note:Bold data in the table indicates the highest score in the current comparison method

5 总结

本文针对现有 MNER 研究中存在的噪声影响和图文语义融合不足的问题,提出了一种多模态语义协同交互的图文联合命名实体识别(ITJNER)模型。以图像描述丰富多模态特征表示和图像语义信息的表达,减少图文交互中有效语义信息的丢失,提出一种将多头跨模态注意力、多头自注意力、门控机制相互耦合的多模态协同交互方法,可以在实现图文语义间有效融合的同时,抑制多模态交互中的不完整或错误的语义信息。实验结果表明,本模型有助于提取图文间的共同语义信息且在图文关联度更高的数据中表现更优,但本模型对于图文关联度较低的数据的准确率仍有待提升。

在未来的工作中,考虑增强模型对图文不相关数据的处理能力,能够排除过滤无关数据噪声对模型的影响,以获得一个更健壮的 NER 模型,同时考虑通过融合知识图谱实现多模态数据的语义表达,并反向推动知识图谱的构建。

参考文献

[1] ESTEVES D, PERES R, LEHMANN J, et al. Named entity recognition in twitter using images and text [C]//International Conference on Web Engineering. Cham, Switzerland: Springer, 2017: 191-199.

[2] ZHANG Q, FU J L, LIU X Y, et al. Adaptive co-attention network for named entity recognition in tweets [C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, Louisiana, USA: AAAI Press, 2018: 5674-5681.

[3] 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. New Orleans, Louisiana, USA: NAACL Press, 2018: 852-860.

[4] 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 (Volume 1: Long Papers). Melbourne, Australia: ACL, 2018: 1990-1999.

[5] 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). Sydney, NSW, Australia: IEEE, 2019: 337-342.

[6] 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, 34(3): 1905-1922.

[7] SUN L, WANG J Q, 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. Palo Alto, California, USA: AAAI Press, 2021, 35(15): 13860-13868.

[8] SUN L, WANG J Q, SU Y D, 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. Barcelona, Spain: ICCL, 2020: 1852-1862.

[9] YU J F, 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. Online: ACL, 2020: 3342-3352.

[10] LIU L P,WANG M L,ZHANG M Z,et al. UAMNer: Uncertainty-aware multimodal named entity recognition in social media posts [J]. Applied Intelligence, 2022,52(4):4109-4125.
[11] TIAN Y,SUN X,YU H F,et al. Hierarchical self-adaptation network for multimodal named entity recognition in social media [J]. Neurocomputing, 2021, 439: 12-21.
[12] ZHENG C M,WU Z W,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.
[13] WU Z W,ZHENG C M,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. Seattle, WA, USA: ACM, 2020: 1038-1046.
[14] ZHANG D,WEI S Z,LI S S,et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, California, USA: AAAI Press,2021,35(16):14347-14355.
[15] CHEN D W,LI Z X,GU B,et al. Multimodal named entity recognition with image attributes and image knowledge [C]//JENSEN C S,LIM E P,YANG D N, et al. Database Systems for Advanced Applications. Cham,Switzerland:Springer,2021:186-201.
[16] CHEN S G,AGUILAR G,NEVES L,et al. A caption is worth a thousand images: Investigating image captions for multimodal named entity recognition [EB/OL]. [2022-03-03]. https://doi.org/10.48550/arXiv.2010.12712.
[17] HUANG F R,LI C Z,GAO B Y,et al. Deep attentive multimodal network representation learning for social media images [J]. ACM Transactions on Internet Technology (TOIT),2021,21(3):1-17.
[18] ZHANG C,YANG Z C,HE X D,et al. Multimodal intelligence: Representation learning, information fusion, and applications [J]. IEEE Journal of Selected Topics in Signal Processing,2020,14(3):478-493.
[19] LAN Z Z,CHEN M D,GOODMAN S,et al. ALBERT: A lite BERT for self-supervised learning of language representations [C]//International Conference on Learning Representations 2020. Addis Ababa, Ethiopia:ICLR,2020:1-17.
[20] GAO S H,CHENG M M,ZHAO K,et al. Res2Net: A new multi-scale backbone architecture [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,43(2):652-662.
[21] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.

Image-Text Joint Named Entity Recognition Method Based on Multi-modal Semantic Interaction

ZHONG Weixing, WANG Hairong, WANG Dong, CHE Miao
(School of Computer Science and Engineering, North Minzu University, Yinchuan, Ningxia, 750021, China)

Abstract: To solve the problem of noise impact and insufficient image-text semantic fusion in existing Multi-modal Named Entity Recognition (MNER) research, this article proposes an Image-Text Joint Named Entity Recognition (ITJNER) model with multi-modal semantic interaction. The ITJNER model adds image description as an additional feature to enrich the multi-modal feature representation. The image description can help filter out the noise introduced from image features and summarize image semantic information in text form. A multi-modal semantic fusion model of multi-modal collaborative interaction is also constructed, which

can enhance the multi-modal information fusion and reduce the semantic deviation of image information. Method experiments were performed on the Twitter-2015 and Twitter-2017 datasets,and the results were analyzed and compared with AdaCAN, UMT, UMGF, Object-AGBAN and other methods. Compared with UMGF method,which showed the optimal result in the above methods,the accuracy,recall and F1 value of this method on Twitter-2017 dataset increased by 0.67%,2.26% and 0.93%,respectively. On the Twitter-2015 dataset,the recall rate increased by 0.19%. The experimental results verify the effectiveness of this method.

Key words:multi-modal named entity recognition;image-text data;multi-modal attention;image description;the semantic integration

责任编辑:梁　晓