多尺度视觉语义增强的多模态命名实体识别方法

王海荣[1,2] 徐 玺[1] 王 彤[1] 陈芳萍[1]

摘 要 为解决多模态命名实体识别 (Multimodal named entity recognition, MNER) 方法研究中存在的图像特征语义缺失和多模态表示语义约束较弱等问题, 提出多尺度视觉语义增强的多模态命名实体识别方法 (Multi-scale visual semantic enhancement for multimodal named entity recognition method, MSVSE). 该方法提取多种视觉特征用于补全图像语义, 挖掘文本特征与多种视觉特征间的语义交互关系, 生成多尺度视觉语义特征并进行融合, 得到多尺度视觉语义增强的多模态文本表示; 使用视觉实体分类器对多尺度视觉语义特征解码, 实现视觉特征的语义一致性约束; 调用多任务标签解码器挖掘多模态文本表示和文本特征的细粒度语义, 通过联合解码解决语义偏差问题, 从而进一步提高命名实体识别准确度. 为验证该方法的有效性, 在 Twitter-2015 和 Twitter-2017 数据集上进行实验, 并与其他 10 种方法进行对比, 该方法的平均 F1 值得到提升.

关键词 多模态命名实体识别, 多任务学习, 多模态融合, Transformer

引用格式 王海荣, 徐玺, 王彤, 陈芳萍. 多尺度视觉语义增强的多模态命名实体识别方法. 自动化学报, 2024, **50**(6): 1234−1245

DOI 10.16383/j.aas.c230573

Multi-scale Visual Semantic Enhancement for Multimodal Named Entity Recognition Method

WANG Hai-Rong[1,2] XU Xi[1] WANG Tong[1] CHEN Fang-Ping[1]

Abstract To address the issues of semantic loss in image features and weak semantic constraints in multimodal representations encountered in the research of multimodal named entity recognition (MNER) methods, multi-scale visual semantic enhancement for multimodal named entity recognition method (MSVSE) is proposed. After supplementing image semantics by extracting multiple visual features, the semantic interaction and feature fusion between text features and various visual features are explored through a multimodal feature fusion module. This process outputs multi-scale visual semantic-enhanced multimodal text representations. The visual entity classifier is used to decode multi-scale visual semantic features to learn the semantic consistency between various visual features. The multi-task decoder is invoked to mine the fine-grained semantic representation in multimodal text repre-sentation and text features, and carry out joint decoding to solve the semantic bias problem, thereby further improving the accuracy of named entity recognition. To verify the effectiveness of the method, experiments were carried out on Twitter-2015 and Twitter-2017 respectively, and compared with other 10 methods. The average F1 values of the MSVSE on the two datasets have increased.

Key words Multimodal named entity recognition (MNER), multi-task learning, multimodal fusion, Transformer

Citation Wang Hai-Rong, Xu Xi, Wang Tong, Chen Fang-Ping. Multi-scale visual semantic enhancement for multimodal named entity recognition method. *Acta Automatica Sinica*, 2024, **50**(6): 1234−1245

收稿日期 2023-09-13 录用日期 2024-02-22
Manuscript received September 13, 2023; accepted February 22, 2024
宁夏自然科学基金 (2023AAC03316), 宁夏回族自治区教育厅高等学校科学研究重点项目 (NYG2022051) 资助
Supported by Natural Science Foundation of Ningxia (2023 AAC03316) and Key Research Project of Education Department of Ningxia Hui Autonomous Region (NYG2022051)
本文责任编委 刘洋
Recommended by Associate Editor LIU Yang
1. 北方民族大学计算机科学与工程学院 银川 750021 2. 北方民族大学图像图形智能处理国家民委重点实验室 银川 750021
1. School of Computer Science and Engineering, North Minzu University, Yinchuan 750021 2. The Key Laboratory of Images & Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021

多模态命名实体识别 (Multimodal named entity recognition, MNER) 通过挖掘文本、图片、音频和视频等多模态数据中的语义特征, 用于辅助多模态信息抽取.

自 2018 年 Moon 等[1] 首次提出多模态命名实体识别方法后, MNER 研究备受关注, 如基于视觉注意力方法[2]、字符−单词−图像特征融合方法[3] 等. 这些方法通过注意力机制和门控机制挖掘图文特征中的语义互补关系, 增强文本特征语义. 但由于文本特征语义层次较低, 挖掘语义互补关系较为

困难, 致使多模态命名实体识别效果不佳. 为了提升文本特征语义, 基于共注意力方法[4]、基于双线性注意力对抗网络方法[5] 和基于密集共注意力方法[6] 等使用双向长短期记忆递归神经网络, 挖掘文本特征中上下文语义, 实现对文本语义的增强后再进行多模态特征融合. 但由于上述方法使用的均是静态文本特征, 无法有效解决图文语义鸿沟问题. 为此, Yu 等[7] 使用双向编码器表征法 (Bidirectional encoder representations from transformers, BERT) 提取动态文本特征, 将其与图像特征进行融合, 从而得到较高质量的多模态文本表示. 目前, 多模态命名实体识别方法研究大多聚焦于多模态特征的深度融合和多模态语义偏差校正 2 个方面.

为实现多模态特征的深度融合, 一些研究者认为挖掘多模态特征间关系对多模态特征的对齐和充分融合是关键点. 如 Xu 等[8] 通过跨模态匹配计算图文相似度, 以确定保留的图像信息, 再进行特征融合来获得最终的跨模态表示; Wang 等[9] 为进一步对齐图文特征, 提出一种挖掘图文特征间的精细化语义关系方法. 此外, 还有一些方法 (如基于统一多模态图融合 (Unified multimodal graph fusion, UMGF) 方法[10]、图文联合命名实体识别方法[11] 和基于分层自适应网络方法[12] 等) 调用多个跨模态注意力机制, 来挖掘模态内部语义关系和模态间语义关系. 以上方法实现了图文特征的充分融合, 但生成的多模态表示中包含了视觉特征中的增益信息, 同时也引入了部分视觉语义噪声, 这导致了多模态语义偏差问题.

为了校正多模态语义偏差, 一些研究者基于多任务学习方法, 提出联合解码策略, 典型的有基于图像–文本对齐的多模态命名实体识别 (Image-text alignments for multimodal named entity recognition, ITA) 方法[13]、具有不确定性感知的多模态命名实体识别方法 (Uncertainty aware multimodal named entity recognition, UAMNer)[14]、基于多任务学习的多模态命名实体识别方法[15]、场景图驱动的多粒度多任务学习的多模态命名实体识别方法 (Scene graph driven multi-granularity multi-task learning for multimodal named entity recognition, M3S)[16] 等. 这类方法通过消除多模态特征和文本特征的预测结果差异, 来解决图文语义冲突等因素导致的视觉偏差问题, 但是没有直接对视觉特征进行优化. 为此, Chen 等[17] 使用动态门控机制优化视觉特征, 并与多模态关系抽取任务联合训练, 从而得到通用性较强的多模态特征; Jia 等[18] 构建了细粒度视觉特征查询任务来增强图像语义理解; Sun 等[19−20] 相继提出全局级、特征级的图文关系预测方法, 对视觉特征过滤和筛选后, 与文本特征进行融合; Xu 等[21] 将图文关系表示为二进制, 当图文关系表示为 0, 则丢弃图像, 仅使用文本数据进行信息抽取; Zhao 等[22] 通过图文对间的语义关系, 收集与当前图文对最相关的图像信息, 来丰富图像语义; Zhou 等[23] 采用变分自编码器 (Variational auto-encoders, VAE), 对图文数据进行统一表示, 以消除图文特征间的语义鸿沟, 并促进多模态特征语义融合.

综上所述, 现有 MNER 方法基本实现图文特征融合, 但仍然存在以下 2 个问题: 1) 主要关注单尺度视觉特征与文本特征间语义交互, 而较少关注单尺度视觉特征中存在的语义缺失问题, 也较少关注多尺度视觉特征与文本特征的语义交互关系的挖掘方法研究. 受数据集规模、领域以及训练目标任务的影响, 当在社交领域 MNER 数据集中使用视觉模型来表示视觉特征时, 视觉语义将被进一步削弱. 2) 仅在图文关系和文本特征上约束语义表示, 而未对视觉特征进行语义约束, 会带来语义约束较弱问题.

为此, 本文提出一种多尺度视觉语义增强的多模态命名实体识别方法 (Multi-scale visual semantic enhancement for multimodal named entity recognition method, MSVSE). 该方法通过挖掘文本特征和多种视觉特征间的多尺度语义交互关系, 以补全图像语义, 得到多尺度视觉语义特征, 并深度融合图文特征, 得到多尺度视觉语义增强的多模态表示. 多模态表示由多模态视觉表示和多模态文本表示组成. 该方法使用视觉实体分类器对多尺度视觉语义特征进行监督学习, 实现对视觉特征的语义一致性约束; 调用多任务标签解码器挖掘多模态文本表示和文本特征的细粒度语义表示, 通过联合解码来解决语义偏差问题, 进而增强多模态文本表示的通用性, 从而进一步提高命名实体识别准确度.

1 MSVSE 方法模型

MSVSE 方法调用多种视觉模型提取多尺度视觉特征, 协同表示图像语义; 通过多模态特征融合模块挖掘文本特征和多尺度视觉特征的语义交互关系, 生成多尺度视觉语义特征, 进行特征融合后, 得到多尺度视觉语义增强的多模态文本表示.

该方法使用视觉实体分类器对多尺度视觉语义特征进行解码, 以实现多尺度视觉语义特征的语义一致性约束, 从而过滤视觉语义噪声, 并消除图文语义冲突. 使用聚合命名实体识别、实体边界检测、实体类别检测和实体存在性检测 4 个任务来挖掘多

模态文本表示中的细粒度语义, 从而提高预测特征的语义准确性, 便于条件随机场解码. 进一步使用多任务标签解码器, 对多模态文本表示和文本特征进行联合解码, 以解决语义偏差问题, 从而提高命名实体识别准确性. MSVSE 模型框架如图 1 所示.

2 多模态特征提取

对于输入的图文数据, MSVSE 方法的首要工作是使用语言或视觉预训练模型, 提取文本特征和多尺度视觉特征, 得到完备的图文语义表示, 主要包含文本特征提取和多尺度视觉特征提取 2 个模块.

2.1 文本特征提取

对输入句子进行转换, 得到单词嵌入 $S = \{[\text{CLS}], S_1, S_2, \cdots, S_{n-1}, [\text{SEP}]\}$, 调用 BERT, 提取文本特征 $H^s = \{H_0^s, H_1^s, \cdots, H_{n-1}^s\}$, 可表示为:

$$H^s = \text{BERT}(S),\ H^s \in \mathbf{R}^{n\times d} \tag{1}$$

式中, n 为句子长度, d 为特征编码维度.

2.2 多尺度视觉特征提取

分别调用预训练视觉模型 Mask-RCNN[24]、图像−字幕 (Image-caption, IC) 模型[25]、残差神经网络 (Residual neural network, ResNet) 模型[26], 提取视觉标签、图像描述和区域视觉特征, 协同表示图像语义, 进而解决单尺度视觉特征中图像语义的缺失问题. 多尺度视觉特征分别表示为:

$$G = \text{ResNet}(I) \tag{2}$$

$$C = \text{Mask-RCNN}(I) \tag{3}$$

$$D = \text{IC}(I) \tag{4}$$

式中, I 是图像向量, 区域视觉特征 G 包含 $\tilde{g}$ 个区域特征的集合, 视觉标签 C 包含 $\tilde{o}$ 个单词的集合, 图像描述 D 是一个包含 $\tilde{d}$ 个单词的句子.

3 多模态特征融合

多模态特征融合模块依次对多尺度视觉特征进行表示、过滤和动态映射等操作, 以生成多尺度视觉语义特征、多尺度视觉语义前缀. 调用 BERT 模型, 对文本特征、图像描述和多尺度视觉语义前缀进行联合编码, 得到多尺度视觉语义增强的多模态文本表示和多模态视觉表示. 多模态特征融合过程如图 2 所示.

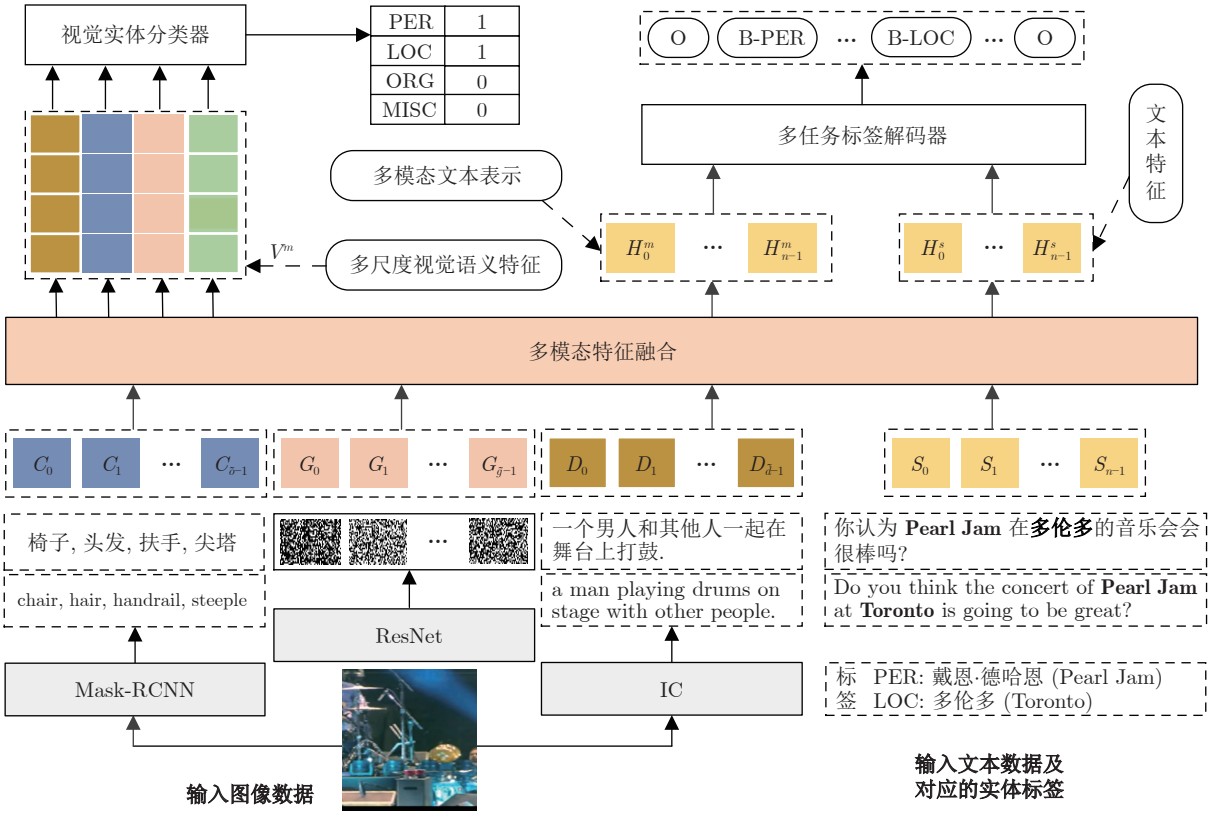

图 1 MSVSE 模型框架

Fig. 1 The framework of MSVSE model

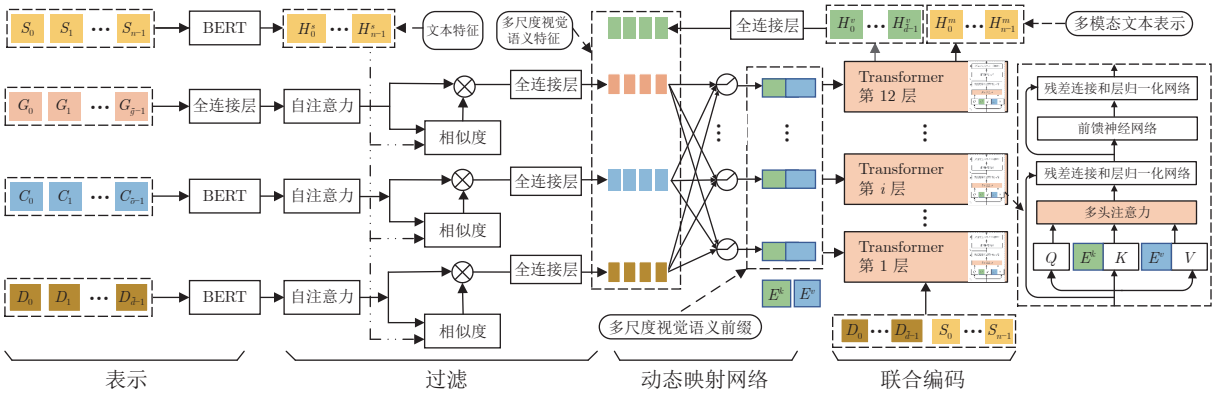

图 2　多模态特征融合模块

Fig. 2　The multimodal feature fusion module

3.1　表示

通过线性层对区域视觉特征进行投影, 使其与文本特征的特征维度一致, 并使用 BERT 分别对视觉标签和图像描述进行特征表示. 多尺度视觉特征表示如下:

$$V_{g'} = W_g G + b_g,\ V_{g'} \in \mathbf{R}^{\tilde{g}\times d} \tag{5}$$

$$V_{o'} = \mathrm{BERT}(C),\ V_{o'} \in \mathbf{R}^{\tilde{o}\times d} \tag{6}$$

$$V_{d'} = \mathrm{BERT}(D),\ V_{d'} \in \mathbf{R}^{\tilde{d}\times d} \tag{7}$$

式中, W_g 和 b_g 分别表示线性投影层的权重矩阵和偏置参数, d 为视觉特征维度. 多尺度视觉特征集合记为 $\{V_{g'}, V_{o'}, V_{d'}\}$.

3.2　过滤

多尺度视觉特征相较于单尺度视觉特征具有更丰富的潜在语义信息, 但也包含更多的视觉噪声. 因此, 有必要使用自注意力机制和相似度模型来挖掘多尺度视觉特征中的显著对象, 并过滤视觉噪声, 从而得到多尺度视觉语义特征.

当使用 V_i 表示多尺度视觉特征集合中的一种单尺度视觉特征, 过滤计算如下:

$$V_i^a = \mathrm{softmax}\left(\frac{[W_1 V_i]^{\mathrm{T}}[W_2 V_i]}{\sqrt{d}}\right)[W_3 V_i]^{\mathrm{T}} \tag{8}$$

$$V_i^s = \frac{H^s \cdot V_i^a}{\parallel H^s \parallel \cdot \parallel V_i^a \parallel} \tag{9}$$

$$V_i^m = w_i(V_i^s \otimes V_i^a) + b_i \tag{10}$$

式中, $V_i \in \{V_{g'}, V_{o'}, V_{d'}\}$, W_1、W_2、W_3 分别为自注意力机制的内部 query、key、value 向量投影层的权重矩阵, w_i、b_i 分别表示用于多尺度视觉特征压缩的全连接网络的权重矩阵、偏置参数.

此外, 本文还调用视觉实体分类器进行语义约束, 以增强多尺度视觉语义特征 V^m 的语义准确性和约束视觉模态语义的一致性 (详见第 4.1 节).

过滤后生成的多尺度视觉语义特征 $V^m = \{V_{g'}^m, V_{o'}^m, V_{d'}^m\}$, $V_i^m \in \mathbf{R}^{z\times d}$, $i \in \{g', o', d'\}$, z 表示压缩后的特征数量.

3.3　动态映射网络

构建动态映射网络, 针对 12 个 Transformer 编码层中注意力的不同语义需求, 动态地对多尺度视觉语义特征进行映射和过滤, 以便生成 Transformer 各编码层所需的多尺度视觉语义特征, 进而辅助多模态特征的语义融合.

将全连接神经网络多尺度视觉语义特征 V^m 中的每一个特征 V_i^m 投影到视觉前缀空间, 计算公式如下:

$$E_i^p = W_i^p V_i^m + b_i^p \tag{11}$$

式中, W_i^p 和 b_i^p 分别为 3 个全连接神经网络的权重参数和偏置参数, $E_i^p \in \mathbf{R}^{z\times 2\times h\times d_u}$, $2\times h\times d_u$ 为视觉前缀空间的特征维度, h 为多头注意力机制中的注意力头数量, d_u 为多头注意力机制的 key 和 value 的编码维度.

使用 12 层门控网络组成的动态映射网络对多尺度视觉语义特征进行融合, 将 E^p 映射为多尺度视觉语义前缀. 第 j 层门控网络表示为:

$$\mathrm{gate}_j(\cdot) = \mathrm{softmax}(\mathrm{ReLU}(w_j'(\cdot) + b_j')) \tag{12}$$

式中, w_j' 和 b_j' 为第 j 个门控网络中全连接神经网络的权重矩阵和偏置参数.

依次将 E_i^p 输入第 j 层门控网络, 为 Transformer 编码层中每个注意力头生成门控信号后, 将该门控信号与 E_i^p 进行向量乘法运算, 得到 E_i^p 的视觉前缀特征. 对第 j 层的所有视觉前缀特征进行求和, 得到第 j 层中语义聚合的视觉前缀 E_j^{kv}, 调用 split

函数切分 E_j^{kv}, 得到第 j 个视觉前缀的 2 个值, 分别对应于多头注意力中的 key 和 value, 计算如下:

$$E_j^{kv} = \sum_{i=\{g',\, o',\, d'\}} E_i^p \cdot \mathrm{gate}_j(E_i^p) \tag{13}$$

$$\left(E_j^k,\ E_j^v\right) = \mathrm{split}(E_j^{kv}) \tag{14}$$

式中, $E_j^{kv} \in \mathbf{R}^{z\times 2\times d_u\times h}$, $E_j^k \in \mathbf{R}^{h\times z\times d_u}$, $E_j^v \in \mathbf{R}^{h\times z\times d_u}$, $E^{kv} \in \mathbf{R}^{l\times z\times 2\times d_u\times h}$, $E^k \in \mathbf{R}^{l\times h\times z\times d_u}$, $E^v \in \mathbf{R}^{l\times h\times z\times d_u}$. 其中$l$ 为 Transformer 编码层数.

3.4 联合编码

为引导文本特征−多尺度视觉特征语义融合, 将多尺度视觉语义前缀作为线索, 使用 BERT 对视觉描述和文本进行联合编码, 得到多尺度视觉语义增强的多模态视觉特征和多模态文本特征, 表示如下:

$$(H^v,\ H^m) = \mathrm{BERT}([D;S],\ (E^k,\ E^v)) \tag{15}$$

式中, BERT 由 12 个 Transformer 编码层组成, 多尺度视觉语义前缀 $(E^k,\ E^v)$ 将按层更新 Transformer 编码层中的多头注意力权重. 具有多尺度视觉前缀的多头注意力计算公式如下:

$$\mathrm{MHA}_j = \mathrm{softmax}\left(\frac{Q_j[E_j^k;K_j]}{\sqrt{d}}\right)[E_j^v;V_j] \tag{16}$$

式中, MHA_j 为第 j 个编码层的多头注意力机制, Q_j、K_j、V_j 分别为多头注意力的 query、key 和 value 向量. $[E_j^k;K_j]$ 表示将多尺度视觉语义前缀中第 j 个 key 与第 j 个编码层中多头注意力机制 key 进行拼接, 作为新的 key; $[E_j^v;V_j]$ 表示将多尺度视觉语义前缀中第 j 个 value 与第 j 个编码层中多头注意力机制 value 进行拼接, 作为新的 value, 用于更新多头注意力机制的权重, 促进多模态语义关系挖掘和特征融合.

4 多任务协同处理

多任务协同处理由视觉实体分类器和多任务标签解码器 2 个部分组成, 该算法通过视觉实体分类器对多尺度视觉语义特征进行解码, 实现视觉语义一致性表示. 通过多任务标签解码器对多模态文本表示和文本特征进行细粒度语义挖掘和解码, 以获得最优标签.

4.1 视觉实体分类器

本文使用 BIO (Begin, inside, outside) 实体标注法定义实体标签, 包括人名实体 (Person, PER) 的开始字符 (Begin person, B-PER), 人名实体的内部字符 (Inside person, I-PER); 地名实体 (Location, LOC) 的开始字符 (Begin location, B-LOC), 地名实体的内部字符; 机构名实体 (Organization, ORG) 的开始字符, 机构名实体的内部字符; 非实体 (Out-side, O); 杂项 (Miscellaneous, MISC). 基于图像语义和文本语义的全局一致性, 将命名实体标签转化为多尺度视觉语义特征的全局视觉实体软标签, 转换规则为设视觉实体集合为 NE = [PER, LOC, ORG, MISC], 分别对应人名实体识别的 F1 值、地名实体识别的 F1 值、机构名实体识别的 F1 值和 MISC 识别的 F1 值. 视觉标签序列 $L^E \in \mathbf{R}^{4\times 1}$, $L^E[i] \in [0,\ 1]$, 其中每个值分别表示 NE 中的对应实体是否存在, 如对于 PER, 当文本命名实体标签 Y 中包含 B-PER 或 I-PER 时, $L^E[0] = 1$; 否则, $L^E[0] = 0$. 视觉标签序列可表示为:

$$L^E[i] = \begin{cases} 1,\ NE[i] \in Y \\ 0,\ NE[i] \notin Y \end{cases} \quad i \in [0,\ 1,\ 2,\ 3] \tag{17}$$

使用共享的多层感知机 (Multi-layer perceptron, MLP) 对视觉特征进行分类, 调用交叉熵函数 $\mathrm{CE}(p,\ q) = -\sum_{i=1}^{n}(p(x_i)\ln q(x_i))$, 其中 p 和 q 为概率分布函数, 计算损失如下:

$$\mathcal{L}_{\mathrm{VE}} = \sum_{v\in\{V_{g'}^m,\ V_{o'}^m,\ V_{d'}^m,\ H^v\}} \mathrm{CE}(\mathrm{MLP}(v),\ L^E) \tag{18}$$

当 $\mathcal{L}_{\mathrm{VE}}$ 取得最小值时, 说明多尺度视觉语义特征 $\{V_{g'}^m,\ V_{o'}^m,\ V_{d'}^m,\ H^v\}$ 中的每个值均表示与视觉标签序列最相似的标签语义, 即各个视觉特征间具有语义一致性.

4.2 多任务标签解码器

根据实体边界检测、实体类别检测、实体存在性检测任务与命名实体任务间的标签语义转换关系, 构建了 T_2、T_5、T_7 和 T_{11} 四个投影矩阵, 将这四个子任务的预测特征投影到命名实体识别任务预测向量空间, 共同挖掘特征中的细粒度语义, 进而增强预测特征语义的准确度, 便于调用条件随机场进行解码. 多任务标签解码器如图 3 所示.

调用线性层对输入特征进行预测, 得到四个子任务的预测向量:

$$H_p = W_p H + b_p \tag{19}$$

式中, $H \in \mathbf{R}^{n\times d}$ 为输入特征, $H_p \in \mathbf{R}^{n\times p}$, $p \in [2,\ 5,\ 7,\ 11]$, W_p 为线性层的权重矩阵, b_p 为线性层的偏置参数.

通过投影矩阵 T_2、T_5、T_7 和 T_{11}, 将对应的 H_2、H_5、H_7 和 H_{11} 转换到命名实体识别任务的预测空间, 使用向量加法运算得到最终的预测向量:

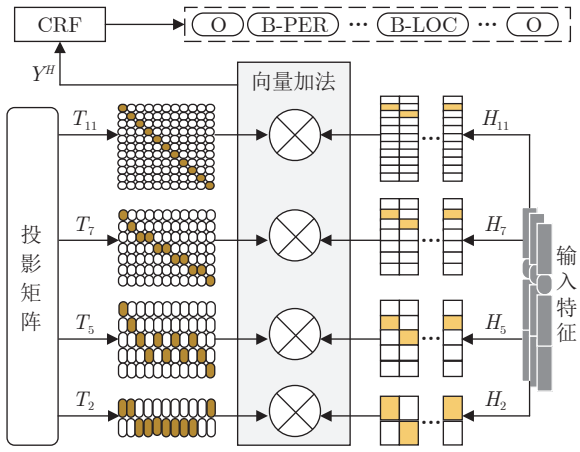

图 3　多任务标签解码器

Fig. 3　The multi-task label decoder

$$Y^H = H_2 \otimes T_2 + H_5 \otimes T_5 + H_7 \otimes T_7 + H_{11} \otimes T_{11} \quad (20)$$

考虑到标签间的依赖关系, 利用条件随机场 (Conditional random field, CRF) 来标记 Y^H. 损失表示如下:

$$\mathcal{L}_{\mathrm{MTD}}^{H} = \mathrm{CRF}(Y^H,\ Y) \quad (21)$$

式中, Y^H 为预测标签, Y 为真实标签.

调用多任务解码对多模态文本表示 H^m 和文本特征 H^s 进行联合解码, 得到预测标签和损失表示如下:

$$\mathcal{L}_{\mathrm{MTD}}^{m},\ Y^m = \mathrm{MTD}\,(H^m,\ Y) \quad (22)$$

$$\mathcal{L}_{\mathrm{MTD}}^{s},\ Y^s = \mathrm{MTD}\,(H^s,\ Y) \quad (23)$$

式中, MTD(·) 表示式 (19) ~ (21) 的运算集合, 代表多任务标签解码器.

通过最小化预测标签序列 Y^m 和 Y^s 的结果差异, 学习文本特征和多模态文本表示的语义一致性, 以解决语义偏差的问题. 计算如下:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{y \in [Y^s,\ Y^m]} p(y|Y^s) \ln p(y|Y^m) \quad (24)$$

MSVSE 方法的预测标签为 Y^m, MSVSE 方法的最终损失函数为:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MTD}}^{s} + \mathcal{L}_{\mathrm{MTD}}^{m} + \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{VE}} \quad (25)$$

式中, $\mathcal{L}_{\mathrm{MTD}}^{s}$ 和 $\mathcal{L}_{\mathrm{MTD}}^{m}$ 分别表示文本特征和多模态文本表示的预测情况, $\mathcal{L}_{\mathrm{KL}}$ 和 $\mathcal{L}_{\mathrm{VE}}$ 分别表示文本模态和视觉模态中特征语义的一致性情况.

5　方法验证及结果分析

5.1　实验设计

为了验证 MSVSE 方法的有效性, 使用 pytorch 技术搭建实验环境, 在 Twitter-2015、Twitter-2017 公共多模态命名实体识别数据集上进行实验, 使用评价指标 F1 值对 MSVSE 方法进行性能评估. 使用 AdamW 优化器调整模型参数. 训练轮数设置为 30, 批次大小设置为 32, 学习率设置为 3×10^{-5}.

5.2　对比方法

为了评估本文方法的有效性, 选择文本特征使用 BERT 提取的 10 种 MNER 模型作为基线模型.

1) 基于多模态小双向编码器表征法 (Multimodal small bidirectional encoder representations from transformers, MSB) 的多模态命名实体识别方法[3] 模型采用小 BERT 语言模型对图文特征进行联合编码, 以生成多模态文本表示. 图像特征是包含 5 个单词的图像分类标签.

2) 基于统一多模态 Transformer 和实体跨度检测改进的多模态命名实体识别方法 (Improving multimodal named entity recognition via entity span detection with unified multimodal transformer, UMT)[7] 模型采用 3 个跨模态注意力机制挖掘多模态特征间交互作用, 生成多模态文本表示. 引入边界检测任务识别文本特征中的边界语义, 辅助多模态命名实体识别.

3) 基于通用匹配与对齐框架的多模态命名实体识别方法 (A general matching and alignment framework for multimodal named entity recognition, MAF)[8] 模型采用跨模态注意力机制来挖掘图文特征的一一对应关系, 实现特征对齐后再生成多模态文本表示.

4) UMGF 模型构建了区域视觉特征和文本特征的图结构表示, 通过多层具有注意力机制的跨模态门控机制来聚合图文语义, 生成多模态文本表示. MAF 和 UMGF 均采用区域视觉特征.

5) UAMNer 模型在文本特征和多模态文本表示上分别构建命名实体识别任务. 引入贝叶斯神经网络计算文本特征中预测标签的不确定性. 使用不确定性较高的文本特征命名实体识别标签替换多模态文本表示的标签, 以得到更为准确的预测标签.

6) M3S 模型提取图像中的场景图特征, 使用图神经网络来聚合图文语义, 生成多模态文本表示.

7) 基于分层视觉前缀融合网络的多模态命名实体识别 (Hierarchical visual prefix fusion network for multimodal entity extraction, HvpNet)[17] 模型将层次视觉特征映射为前缀, 调用 BERT 对文本进行编码, 以生成多模态文本表示.

8) 基于查询的多模态命名实体识别 (Multimod-

al named entity recognition with query grounding, MNER-QG) 方法[18] 通过人工标注视觉特征的细粒度标签, 利用视觉查询任务来优化层次视觉特征, 使其语义表述更为准确; 再采用机器阅读理解的方法融合图文特征.

9) 基于关系增强图卷积网络的多模态命名实体识别 (Relation-enhanced graph convolutional network for multimodal named entity recognition, RGCN)[22] 模型通过检索得到数据集中与当前图文对相关的多张图片, 以补充视觉语义, 并利用图神经网络和跨模态注意力机制来融合图文特征.

10) VAE 模型使用文本 VAE 和图像 VAE 构建多模态变分自动编码器, 以提取图文特征. 基于图文特征的均值和高斯分布等信息引导图文特征语义的融合, 得到多模态文本表示.

5.3 方法性能评价

与 10 种 MNER 模型进行对比分析, 实验结果如表 1 所示. 表 1 中, HvpNet 仅包含多模态命名实体识别的实验复现结果, –HvpNet 为 MSVSE 与 HvpNet 的性能差值.

由表 1 可知, 与使用图文联合编码实现图文特征融合的 MSB 模型相比, MSVSE 在 2 个数据集上的 F1 值分别提升了 1.64%、3.02%, 可能的原因是 MSB 仅使用了一种视觉特征, 而 MSVSE 既调用了多尺度视觉特征协同表示图像语义, 也利用视觉实体分类器对多尺度视觉特征进行监督学习, 进而得到了更为丰富和准确的视觉语义, 解决了视觉语义缺失问题.

与使用跨模态注意力机制实现图文特征融合的 MAF、UMGF 等模型相比, MSVSE 在 2 个数据集上的 F1 值分别平均提升了 0.96%、1.46%, 表明相比于堆叠多个跨模态注意力的多模态特征融合模型, 使用多尺度视觉语义前缀优化 BERT 语言模型能更充分融合图文特征, 从而得到高质量的多模态表示. 此外, 与使用图神经网络的 M3S 相比, MSVSE 在 2 个数据集上均取得了良好效果, 再次验证了 MSVSE 中多模态特征融合方法的高效性.

与使用视觉前缀进行图文特征融合的基准模型 HvpNet 相比, 如表 1 最后一行所示, MSVSE 在 2 个数据集上的多个指标取得了较好性能, 其原因在于多尺度视觉语义前缀相比单一的层次视觉特征含有更为准确和丰富的视觉语义.

由对比实验可知, 多任务模型 (如 UMT、UAMNer、MNER-QG 和 VAE) 性能优于单任务模型 (如 MSB、MAF 和 UMGF), 这是因为多任务模型解决了视觉偏差问题. 本文方法通过标签对比损失整合了这种能力, 并且通过多任务标签解码器来增强 CRF 的解码能力. 相比于简单调用 CRF 或 softmax 作为解码器的多任务模型, 在 2 个数据集上, MSVSE 的 F1 值分别平均提升了 0.98%、1.38%.

直接增强或衰减视觉特征对优化视觉语义是有效的, 如 MNER-QG 和 RGCN. MNER-QG 方法通过人工标注视觉特征的细粒度视觉语义标签, 以确保层次视觉特征的语义更为准确, 但人工成本较高. RGCN 方法通过检索得到数据集上与图文对相关的多张图片, 以补充视觉语义, 但在图片检索和图像特征融合过程中可能存在级联误差. 与 RGCN 相比, MSVSE 在 2 个数据集上的 F1 值分别提升了 0.11%、0.23%. 这是因为 MSVSE 方法中构建了

表 1 数据集上方法性能比较 (%)

Table 1 Performance comparison of method on dataset (%)

方法	Twitter-2015					Twitter-2017				
	PER	LOC	ORG	MISC	F1	PER	LOC	ORG	MISC	F1
MSB	86.44	77.16	52.91	36.05	73.47	—	—	—	—	84.32
MAF	84.67	81.18	63.35	41.82	73.42	91.51	85.80	85.10	68.79	86.25
UMGF	84.26	**83.17**	62.45	42.42	74.85	91.92	85.22	83.13	69.83	85.51
M3S	86.05	81.32	62.97	41.36	75.03	92.73	84.81	82.49	69.53	86.06
UMT	85.24	81.58	63.03	39.45	73.41	91.56	84.73	82.24	70.10	85.31
UAMNer	84.95	81.28	61.41	38.34	73.10	90.49	81.52	82.09	64.32	84.90
VAE	85.82	81.56	63.20	**43.67**	75.07	91.96	81.89	84.13	**74.07**	86.37
MNER-QG	85.68	81.42	63.62	41.53	74.94	93.17	86.02	84.64	71.83	87.25
RGCN	86.36	82.08	60.78	41.56	75.00	92.86	**86.10**	84.05	72.38	87.11
HvpNet	85.74	81.78	61.92	40.81	74.33	92.28	84.81	84.37	65.20	85.80
MSVSE	**86.72**	81.63	**64.08**	38.91	**75.11**	**93.24**	85.96	**85.22**	70.00	**87.34**
–HvpNet	0.98	–0.15	2.16	–1.90	0.78	0.96	1.15	0.85	4.80	1.54

视觉实体分类器, 它基于图文对全局语义一致性的假设, 将文本标签迁移转化为图像特征软标签, 减少了人力成本, 避免了语义传递误差, 从而学习了多尺度视觉特征中的模态不变性.

5.4　消融实验

为了验证 MSVSE 模型中各组件的有效性, 在 Twitter-2015、Twitter-2017 数据集上进行消融实验, 以评估自注意力机制、相似度、多任务标签解码器和视觉实体分类器对模型性能的影响. 实验结果如表 2 所示, 其中“w/o”表示从 MSVSE 网络中去除对应的模型结构.

由表 2 可知, w/o 自注意力机制、w/o 相似度模型的性能远低于 MSVSE, 表明自注意力机制挖掘到了视觉特征中的显著对象, 并通过视觉特征与文本特征的语义相似关系, 过滤了无关视觉特征或视觉噪声, 从而增强了视觉语义的准确性.

在 2 个数据集上, w/o 多任务标签解码器的 F1 值分别下降了 0.42%、0.20%, 表明多任务标签解码器能挖掘多模态表征中的实体存在性、实体边界、实体类属等细粒度语义来帮助实体识别. w/o 视觉实体分类器的 F1 值分别下降了 0.32%、0.42%, 可能的原因是通过约束多尺度视觉语义特征语义一致性, 有益于增强多模态表示的通用性, 进而提升实体识别性能.

为了探究联合编码时视觉特征对模型性能的影响, 设置了 4 组对照实验, 分别是文本、视觉标签加文本、视觉标签加图像描述加文本、图像描述加文本 (MSVSE), 在 2 个数据集上的实验结果如表 3 所示. 其中“✓”表示 MSVSE 使用了对应的文本特征或视觉特征, “—”表示 MSVSE 没有使用对应的特征.

由表 3 可知, 融合了图像描述或视觉标签后, 模型性能有了进一步提升.

为了探究多尺度视觉语义前缀中不同视觉特征的重要性, 在 2 个数据集上进行了区域视觉特征、区域视觉特征加视觉标签、区域视觉特征加图像描述、区域视觉特征加图像描述加视觉标签 (MSVSE) 四组对比实验, 实验结果如表 4 所示. 其中“✓”表示 MSVSE 使用了对应的视觉特征, “—”表示 MSVSE 没有使用对应特征.

由表 4 可知, 不同视觉特征均有不同程度的语义丢失. 因此, 融合了多种视觉特征的语义信息能得到更为准确和更全面的视觉语义, 以生成更高质量的多尺度视觉语义前缀, 从而提升模型性能.

5.5　视觉实体分类性能评价

为了验证不同尺度视觉特征对 MNER 效果的影响, 使用评价指标 F1 值作为视觉实体分类任务的评估指标, 分别在 Twitter-2015、Twitter-2017 数据集上进行了视觉实体分类任务的实验, 用以验证本文采用多尺度视觉特征协同表示的积极作用. 2 个数据集上的实验结果分别如图 4、图 5 所示.

图 4、图 5 中横坐标表示 3 种单尺度视觉特征

表 2　模型结构消融实验 (%)

Table 2　Structural ablation experiments for the model (%)

方法	Twitter-2015					Twitter-2017				
	PER	LOC	ORG	MISC	F1	PER	LOC	ORG	MISC	F1
MSVSE	**86.72**	81.63	**64.08**	38.91	**75.11**	93.24	85.96	85.22	70.00	**87.34**
w/o 自注意力机制	86.49	81.20	63.21	**41.56**	74.83	93.05	86.52	84.37	67.34	86.79
w/o 相似度	86.33	81.59	63.15	40.84	74.91	92.94	**86.59**	84.07	68.24	86.75
w/o 自注意力机制加相似度	86.80	81.38	63.32	39.62	74.67	92.97	85.87	84.41	67.96	86.67
w/o 多任务标签解码器	86.49	81.78	62.68	37.60	74.69	92.98	84.83	85.02	**71.66**	87.14
w/o 视觉实体分类器	86.52	81.64	63.06	39.89	74.79	**93.37**	84.83	**85.82**	66.24	86.92

表 3　联合编码器中视觉特征消融实验 (%)

Table 3　Visual feature ablation experiments in the joint encoder (%)

文本	视觉标签	图像描述	Twitter-2015					Twitter-2017				
			PER	LOC	ORG	MISC	F1	PER	LOC	ORG	MISC	F1
✓	—	✓	86.72	81.63	**64.08**	38.91	**75.11**	93.24	85.96	**85.22**	70.00	87.34
✓	—	—	86.76	81.68	61.21	**39.46**	74.73	92.95	**86.20**	84.60	70.82	87.11
✓	✓	—	**86.87**	81.74	63.72	37.80	74.87	93.03	85.71	84.43	**71.71**	87.16
✓	✓	✓	86.51	**81.85**	62.20	38.36	74.72	**93.73**	85.96	84.62	70.97	**87.38**

表 4 多尺度视觉语义前缀中视觉特征消融实验 (%)

Table 4 Visual feature ablation experiments in multi-scale visual semantic prefixes (%)

区域视觉特征	视觉标签	图像描述	Twitter-2015					Twitter-2017				
			PER	LOC	ORG	MISC	F1	PER	LOC	ORG	MISC	F1
✓	✓	✓	86.72	81.63	**64.08**	**38.91**	**75.11**	**93.24**	85.96	85.22	**70.00**	**87.34**
✓	—	—	86.25	**81.93**	63.99	38.23	74.76	93.16	84.83	85.47	69.10	87.13
✓	✓	—	86.56	81.60	64.01	38.59	74.93	93.02	85.79	**85.97**	68.67	87.28
✓	—	✓	**86.87**	81.79	63.36	38.68	74.98	92.94	**86.52**	85.14	68.94	87.14

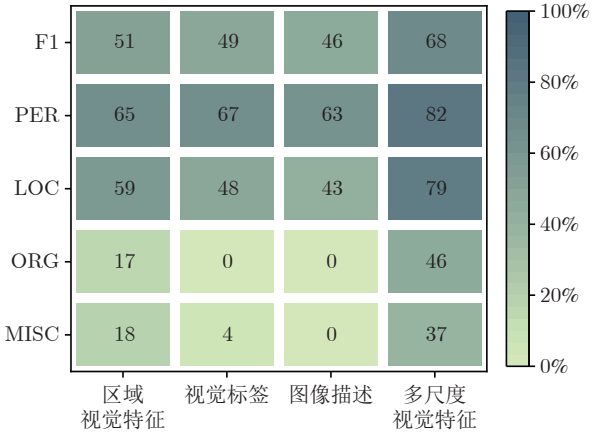

图 4 在 Twitter-2015 上的视觉实体分类性能比较

Fig. 4 Performance comparison of visual entity classification on Twitter-2015

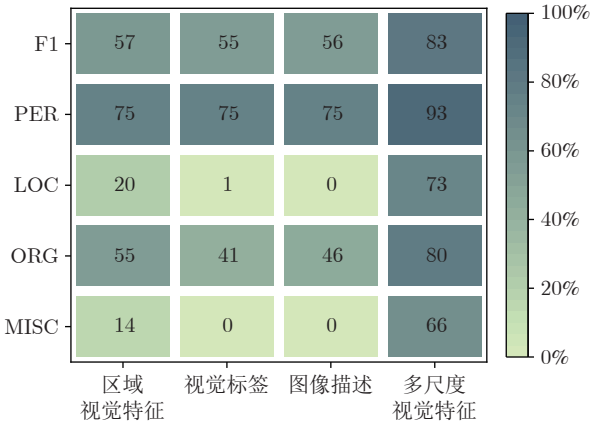

图 5 在 Twitter-2017 上的视觉实体分类性能比较

Fig. 5 Performance comparison of visual entity classification on Twitter-2017

和多尺度视觉特征, 纵坐标表示 5 个评价指标, 用来评价特定视觉特征下视觉实体语义表示的性能.

由图 4、图 5 可以看出, 多尺度视觉特征在 5 个评价指标上均表现最佳, 说明 MSVSE 采用的融合多视觉特征协同表示方法可有效地生成语义准确的多模态文本表示.

多尺度视觉特征协同表示方法在 MISC 类实体上的识别效果不佳, 其主要原因是多尺度视觉语义前缀传递的是全局视觉语义, 但处理时仅对图像描述和文本进行了联合编码, 而由于视觉描述中 MISC 类实体的语义评价值为 0, 导致 MSVSE 方法没有融合到细粒度的 MISC 类实体的语义. 这也进一步解释了表 1 中本文方法的 MISC 识别结果比 VAE 低的原因. 虽然 VAE 方法通过视觉自编码器或图片检索来丰富视觉信息, 使得 MISC 类实体的效果较好, 但也带来了 PER、LOC 和 ORG 类实体识别效果不佳问题.

5.6 单尺度视觉特征性能评估

MSVSE 对单尺度视觉特征提取中语义丢失问题的解决方法不仅在多尺度特征提取中是有效的, 而且在仅采用单尺度视觉特征提取的场景中也是有效的. 为验证本文模型在单尺度视觉特征提取效果, 在 Twitter-2015、Twitter-2017 数据集上, 与 MSB、MAF 和 ITA 进行对比, 实验结果如表 5 所示.

表 5 单尺度视觉特征下方法性能对比 (%)

Table 5 Performance comparison of methods under single scale visual feature (%)

方法	单尺度视觉特征	Twitter-2015 F1	Twitter-2017 F1
MAF	区域视觉特征	73.42	86.25
MSB	图像标签	73.47	84.32
ITA	视觉标签	**75.18**	85.67
ITA	5 个视觉描述	75.17	85.75
ITA	光学字符识别	75.01	85.64
MSVSE	only 区域视觉特征	74.84	86.75
MSVSE	only 视觉标签	74.66	87.17
MSVSE	only 视觉描述	74.56	87.23
MSVSE	w/o 视觉前缀	74.89	87.08
MSVSE (本文方法)		75.11	**87.34**

表 5 中, “only 视觉标签”表示 MSVSE 模型仅使用一种视觉特征即视觉标签, 但将其投影为视觉前缀, 以补充图像的全局语义; “w/o 视觉前缀”表示仅使用视觉描述这一种视觉特征, 并且不将其投影为视觉前缀.

由表 1 和表 5 可知, 在仅采用单尺度视觉特征时, 本文 MSVSE 方法性能超过使用同样特征的 MSB、MAF、UMGF、M3S、UMT、UAMNer、VAE 和 ITA. 但在表 5 的 Twitter-2015 数据集上, F1 值略低的原因可能是这些方法将视觉特征模型加入训练过程中, 并对视觉特征进行了优化, 从而达到更好结果. 如 MNER-QG 对局部视觉特征进行了细粒度标注, 用来获取更有效的视觉语义; M3S 提取的视觉场景图特征相比图像标签、视觉描述, 包含了更为全面的视觉实体信息和视觉实体间关系信息; VAE 通过编码器来优化视觉特征; RGCN 利用多模态图文检索方法获取 6 个图像数据来表示视觉语义; ITA 采用 5 个图像描述表示图像语义.

由“w/o 视觉前缀”实验结果可知, 当去除 MSVSE 的视觉前缀时, 在 2 个数据集上的性能均下降. 其原因在于视觉前缀聚合了多尺度视觉特征中的全局语义, 这有利于辅助引导 BERT 模型在联合编码中生成高质量的多模态文本表示; 融合了多种视觉特征的语义信息, 能得到更为准确和全面的视觉语义, 进而生成高质量的多尺度视觉语义前缀, 进一步提升了模型性能.

5.7　参数及时间效率分析

在 Twitter-2015、Twitter-2017 数据集上进行实验, 通过 F1 值评估不同学习率对 MSVSE 模型性能的影响, 实验结果如表 6 所示.

表 6　不同学习率的方法性能对比 (%)
Table 6　Performance comparison of methods under different learning rates (%)

数据集	学习率 ($\times 10^{-5}$)					
	1	2	3	4	5	6
Twitter-2015	73.4	75.0	**75.1**	74.8	74.6	74.5
Twitter-2017	87.1	86.8	87.3	**87.5**	87.2	87.3

由表 6 可知, 在 2 个数据集上, 当学习率分别为 3×10^{-5} 和 4×10^{-5} 时, F1 值取得最优值.

为了进一步验证模型的复杂性, 进行了模型参数量、单轮训练时间和单轮验证时间对比, 实验结果如表 7 所示. 表 7 中除本文 MSVSE 方法外, 其他方法数据来自文献 [27].

由表 7 可以看出, 本文 MSVSE 方法的参数量没有增加, 这是因为本文方法中的多尺度视觉特征提取是独立的, 其与 BERT 共享参数, 而且视觉实体分类器和多任务标签解码器的参数也共享; 与仅使用 CRF 作为解码器的方法相比, 本文 MSVSE 方法仅增加了 4 个线性层用来提取细粒度语义. 与其他方法相比, 本文 MSVSE 方法的参数量最少, 时间效率也优良.

表 7　参数量及时间效率对比
Table 7　Comparison of parameter number and time efficiency

方法	参数量 (MB)	训练时间 (s)	验证时间 (s)
MSB	122.97	**45.80**	**3.31**
UMGF	191.32	314.42	18.73
MAF	136.09	103.39	6.37
ITA	122.97	65.40	4.69
UMT	148.10	156.73	8.59
HvpNet	143.34	70.36	9.34
MSVSE (本文方法)	**119.27**	75.81	7.03

5.8　基于预训练语言模型的 MNER 的性能评估

文献 [27] 研究发现, 采用不同预训练语言模型表示文本语义对多模态命名实体识别方法性能有不同影响. 因此, 本文分别选取 Glove (Global vector)、BERT、BERT-large、XLMR (Cross-lingual language model and robustly optimized bert pre-training approach)、ChatGPT (Chat generative pre-trained transformer) 五种预训练语言模型表示文本特征, 用于评估其所对应的多模态命名实体识别方法 Glove-BiLSTM-CRF[7]、BERT-CRF[7]、BERT-large-CRF、XLMR-CRF[13] 和 Prompting ChatGPT[28] 性能, 实验结果如表 8 所示.

表 8　基于预训练语言模型的 MNER 方法性能对比 (%)
Table 8　Performance comparison of MNER method based on pre-trained language model (%)

方法	Twitter-2015	Twitter-2017
Glove-BiLSTM-CRF	69.15	79.37
BERT-CRF	71.81	83.44
BERT-large-CRF	73.53	86.81
XLMR-CRF	77.37	89.39
Prompting ChatGPT	79.33	91.43
MSVSE	75.11	87.34

由表 8 可以看出, 随着预训练语言模型的演进, 文本语义表示越来越准确, 促使命名实体识别方法性能随之提升. 然而, 针对多模态命名实体识别方法, 使用的预训练语言模型表示文本语义可能产生歧义, 因此, 可以通过多模态特征融合来校正文本特征语义, 进而提升命名实体识别的准确性. 例如, 相比于 BERT 或 BERT-large 预训练语言模型, 本文 MSVSE 方法表现出了较好性能, 但低于采用 XLMR 和 ChatGPT 方法. 其原因是 XLMR 和 Chat-

GPT 预训练语言模型具有复杂的神经网络结构并使用了超大规模的数据进行预训练, 能得到更加准确的文本语义, 因此使用该模型的多模态命名实体识别方法的性能较为突出. 相比采用 XLMR 和 ChatGPT 方法, 本文方法采用 BERT 模型, 存在文本语义误差.

6 结束语

针对现有 MNER 方法存在图像特征语义缺失和多模态表示语义弱约束问题, 提出多尺度视觉语义增强的多模态命名实体识别方法. 该方法通过挖掘文本特征与多尺度视觉特征间的语义交互关系, 以解决图像特征语义缺失的问题. 利用视觉实体分类器监督多尺度视觉语义特征的生成, 实现视觉特征的实体语义一致性约束. 调用多任务标签解码器对多模态文本表示和文本特征进行预测, 以挖掘特征中的细粒度的实体语义, 来增强预测特征的语义准确性, 从而解决多模态语义偏差问题. 在 Twitter-2015、Twitter-2017 数据集上, 将该方法与其他 10 种方法进行对比实验, 实验结果表明, 该方法能较好地识别多模态数据中的命名实体.

本文通过多尺度视觉特征, 获得了较为全面的视觉语义, 但图像描述等视觉特征仍存在视觉噪声或语义描述错误问题. 在未来研究中, 考虑借助多模态预训练模型来增强文本语义理解, 同时尝试调用视觉大模型 BLIP (Bootstrapping language-image pre-training for unified vision-language understanding and generation)、CogView (Cross-modal general view) 表示图像语义, 以便得到更为全面、准确的视觉特征, 进而增强视觉语义理解, 提升多模态文本表示质量. 此外, 考虑结合图文特征对齐技术和标签迁移技术, 实现对视觉特征的多粒度监督学习, 以获取视觉特征中的有益信息.

References

1 Moon S, Neves L, Carvalho V. Multimodal named entity recognition for short social media posts. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA: NAACL Press, 2018. 852−860

2 Lu D, Neves L, Carvalho V, Zhang N, Ji H. Visual attention model for name tagging in multimodal social media. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: 2018. 1990−1999

3 Asgari-Chenaghlu M, Farzinvash M R, Farzinvash L, Balafar M A, Motamed C. CWI: A multimodal deep learning approach for named entity recognition from social media using character, word and image features. *Neural Computing and Applications*, 2022, **34**(3): 1905−1922

4 Zhang Q, Fu J L, Liu X Y, Huang X J. Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI Press, 2018. 5674−5681

5 Zheng C M, Wu Z W, Wang T, Cai Y, Li Q. Object-aware multimodal named entity recognition in social media posts with adversarial learning. *IEEE Transactions on Multimedia*, 2020, **23**: 2520−2532

6 Wu Z W, Zheng C M, Cai Y, Chen J Y, Leung H F, Li Q. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM, 2020. 1038−1046

7 Yu J F, Jiang J, Yang L, Xia R. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event: 2020. 3342−3352

8 Xu B, Huang S Z, Sha C F, Wang H Y. MAF: A general matching and alignment framework for multimodal named entity recognition. In: Proceedings of the 15th ACM International Conference on Web Search and Data Mining. New York, USA: Association for Computing Machinery, 2022. 1215−1223

9 Wang X W, Ye J B, Li Z X, Tian J F, Jiang Y, Yan M, et al. CAT-MNER: Multimodal named entity recognition with knowledge refined cross-modal attention. In: Proceedings of the IEEE International Conference on Multimedia and Exposition. Taipei, China: 2022. 1−6

10 Zhang D, Wei S Z, Li S S, Wu H Q, Zhu Q M, Zhou G D. Multimodal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2021. 14347−14355

11 Zhong Wei-Xing, Wang Hai-Rong, Wang Dong, Che Miao. Image-text joint named entity recognition method based on multimodal semantic interaction. *Guangxi Sciences*, 2022, **29**(4): 681−690
(钟维幸, 王海荣, 王栋, 车淼. 多模态语义协同交互的图文联合命名实体识别方法. 广西科学, 2022, **29**(4): 681−690)

12 Yu T, Sun X, Yu H F, Li Y, Fu K. Hierarchical self-adaptation network for multimodal named entity recognition in social media. *Neurocomputing*, 2021, **439**: 12−21

13 Wang X Y, Gui M, Jiang Y, Jia Z X, Bach N, Wang T, et al. ITA: Image-text alignments for multimodal named entity recognition. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, USA: Association for Computational Linguistics, 2022. 3176−3189

14 Liu L P, Wang M L, Zhang M Z, Qing L B, He X H. UAMNer: Uncertainty aware multimodal named entity recognition in social media posts. *Applied Intelligence*, 2022, **52**(4): 4109−4125

15 Li Xiao-Teng, Zhang Pan-Pan, Gou Zhi-Nan, Gao Kai. Multimodal named entity recognition method based on multi-task learning. *Computer Engineering*, 2023, **49**(4): 114−119
(李晓腾, 张盼盼, 勾智楠, 高凯. 基于多任务学习的多模态命名实体识别方法. 计算机工程, 2023, **49**(4): 114−119)

16 Wang J, Yang Y, Liu K Y, Zhu Z P, Liu X R. M3S: Scene graph driven multi-granularity multi-task learning for multimodal NER. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023, **31**: 111−120

17 Chen X, Zhang N Y, Li L, Yao Y Z, Deng S M, Tan C Q, et al. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In: Proceedings of the Association for Computational Linguistics. Seattle, USA: Association for Computational Linguistics, 2022.

1607−1618

18 Jia M, Shen L, Shen X, Liao L J, Chen M, He X D, et al. MNER-QG: An end-to-end MRC framework for multimodal named entity recognition with query grounding. *AAAI*, 2022, **37**(7): 8032−8040

19 Sun L, Wang J Q, Su Y D, Weng F S, Sun Y X, Zheng Z W, et al. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In: Proceedings of the 28th International Conference on Computational Linguistics. Virtual Event: 2022. 1852−1862

20 Sun L, Wang J Q, Zhang K, Su Y D, Weng F S. RpBERT: A text-image relation propagation-based BERT model for multimodal NER. In: Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Event: 2021. 13860−13868

21 Xu B, Huang S, Du M, Wang H Y, Song H, Sha C F, et al. Different data, different modalities reinforced data splitting for effective multimodal information extraction from social media posts. In: Proceedings of the 29th International Conference on Computational Linguistics. Virtual Event: 2022. 1855−1864

22 Zhao F, Li C H, Wu Z, Xing S Y, Dai X Y. Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal NER. In: Proceedings of the 30th ACM International Conference on Multimedia, Association for Computing Machinery. New York, USA: 2022. 3983−3992

23 Zhou B H, Zhang Y, Song K H, Guo W Y, Zhao G Q, Wang W B, et al. A span-based multimodal variational autoencoder for semi-supervised multimodal named entity recognition. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. 6293−6302

24 He K M, Gkioxari G, Dollár P, Girshick R. Mask-RCNN. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: 2017. 2980−2988

25 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: 2015. 3156−3164

26 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: 2016. 770−778

27 Wang Hai-Rong, Xu Xi, Wang Tong, Jing Bo-Xiang. Research progress of multimodal named entity recognition. *Journal of Zhengzhou University (Engineering Science)*, 2024, **45**(2): 60−71 (王海荣, 徐玺, 王彤, 荆博祥. 多模态命名实体识别方法研究进展. 多模态命名实体识别方法研究进展. 郑州大学学报 (工学版), 2024, **45**(2): 60−71)

28 Li J Y, Li H, Pan Z, Sun D, Wang J H, Zhang W K, et al. Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge. In: Proceedings of the Association for Computational Linguistics. Singapore: 2023. 2787−2802

王海荣 北方民族大学教授. 2015 年获得东北大学博士学位. 主要研究方向为大数据知识工程与智能信息处理. 本文通信作者.
E-mail: wanghr@nun.edu.cn
(**WANG Hai-Rong** Professor at North Minzu University. She received her Ph.D. degree from Northeastern University in 2015. Her research interest covers big data knowledge engineering and intelligent information processing. Corresponding author of this paper.)

徐 玺 北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为多模态信息抽取.
E-mail: 20217403@stu.nmu.edu.cn
(**XU Xi** Master student at the School of Computer Science and Engineering, North Minzu University. His main research interest is multimodal information extraction.)

王 彤 北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为多模态信息抽取.
E-mail: is_wangtong@163.com
(**WANG Tong** Master student at the School of Computer Science and Engineering, North Minzu University. Her main research interest is multimodal information extraction.)

陈芳萍 北方民族大学计算机科学与工程学院硕士研究生. 主要研究方向为多模态信息抽取.
E-mail: 17393213357@163.com
(**CHEN Fang-Ping** Master student at the School of Computer Science and Engineering, North Minzu University. Her main research interest is multimodal information extraction.)