doi:10.3969/j.issn.1673-9833.2025.01.009

基于视觉与文本语义增强的多模态命名实体识别方法

满芳滕[1]，朱艳辉[2]，张志轩[2]，应旭剑[1]，陈 豪[2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；2. 湖南工业大学 轨道交通学院，湖南 株洲 412007）

摘 要：为了解决视觉特征和文本特征融合后存在部分语义缺失从而导致视觉信息对文本信息的补充有较大偏差的问题，提出了一种基于视觉与文本语义增强的多模态命名实体识别方法。融合 BERT 文本特征提取和 CLIP（contrastive language–image pre-training）视觉特征提取方法，设计了基于协同交叉注意力机制的特征交互单元，以增强视觉信息和文本信息之间的语义关系。CLIP 通过对比学习框架进行预训练，优化模型以正确匹配视觉和对应的文本描述，最大化正样本（匹配的视觉－文本对）的相似性，同时最小化负样本（不匹配的视觉－文本对）的相似性。采用通用领域数据集 TWITTER-2015 和 TWITTER-2017 作为实验数据集。实验结果表明，本模型相比传统方法在多模态命名实体识别任务中的准确率、召回率、F_1 值均有显著提升。

关键词：多模态；命名实体识别；特征融合；语义增强

中图分类号：TP391 **文献标志码**：A **文章编号**：1673-9833(2025)01-0064-08

引文格式：满芳滕，朱艳辉，张志轩，等 . 基于视觉与文本语义增强的多模态命名实体识别方法 [J]. 湖南工业大学学报，2025，39(1)：64-71.

A Multi-Modal Named Entity Recognition Method Based on Visual and Textual Semantic Enhancement

MAN Fangteng[1]，ZHU Yanhui[2]，ZHANG Zhixuan[2]，YING Xujian[1]，CHEN Hao[2]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. College of Rail Transit，Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：In view of a solution of the partial semantic loss in the fusion of visual and textual features, which leads to a significant deviation in the supplementation of visual information to textual information, a multimodal named entity recognition method has thus been proposed based on visual and textual semantic enhancement. A feature interaction unit based on collaborative cross attention mechanism is designed for an enhancement of the semantic relationship between visual information and textual information by integrating BERT text feature extraction and CLIP (contrastive language image pre-training) visual feature extraction methods. CLIP pre-trains through a contrastive learning framework to optimize the model for a correct matching of visual and corresponding text descriptions, thus maximizing the similarity of positive samples (matched visual text pairs) while minimizing the similarity of negative samples (mismatched visual text pairs). The general domain datasets TWITTER-2015 and TWITTER-2017 are adopted as experimental datasets in this article. Experimental results show that compared with traditional methods, this model is characterized with a significantly improved accuracy, recall, and F_1 score in multi-modal named entity recognition tasks.

Keywords：multi-modal；named entity recognition；feature fusion；semantic enhancement

收稿日期：2024-03-30

基金项目：国家自然科学基金资助项目（52272347）；湖南省教育厅科学研究基金资助重点项目（22A0408）

作者简介：满芳滕，男，湖南工业大学硕士生，主要研究方向为自然语言处理与命名实体识别，E-mail：1243874203@qq.com

通信作者：朱艳辉，女，湖南工业大学教授，主要研究方向为自然语言处理与知识工程，E-mail：swayhzhu@163.com

1　研究背景

模态 [1] 是一种生物学上的概念，指人体感官条件下事物发生或存在的方式。多模态 [2] 从感官条件上来讲是指客观实体可以被人类器官感知后所呈现的模态形式，从数据层面上来讲，是指文本、图像、音频、视频等各种模态的信息可以共用表示一个实体，它们之间存在着一些关联。随着信息化时代的到来，特别是互联网的迅速发展，网上滋生了大量的文本、图像、音频、视频等多种模态的信息，多种模态的数据之间存在一定的联系，怎样利用这些联系来构建一个高效、快速、易于存储的知识系统成为了一个很有挑战性的课题。

多模态命名实体识别（multimodal named entity recognition，MNER）任务 [3] 是多模态知识图谱 [1] 构建中的关键一环。Google 公司在 2012 年首次提出知识图谱 [4] 的基本概念。2019 年，Wang M. 等 [5] 构建了一个大型多模态知识图谱 Richpedia。多模态命名实体识别是一种结合了文本和其他模态信息（比如图像、音频和视频等）的命名实体识别（NER）方法。该方法的目标是利用各种模式中的其他信息来提高命名实体识别的性能。在传统的命名实体识别中，系统主要依赖于文本信息识别和分类实体，如人名、地名、组织机构名等。然而，有时仅通过文本可能无法准确地识别和分类实体。例如，同一个名字可能对应多个不同的人或物，只有考虑上下文信息，比如图片或者音频内容，才能更准确地进行识别。

多模态命名实体识别试图解决这个问题，其通过结合文本和其他模式的信息，以获得更全面的上下文理解。例如，结合图像和文本信息可以帮助系统更准确地识别和分类实体，因为图像可以提供额外的上下文信息，如场景、对象、行为等。总之，多模态命名实体识别是一种充分利用各种模式信息，以提高命名实体识别的精度和鲁棒性的方法。

多模态命名实体识别任务最早是由 S. Moon 等 [6] 提出，来自于实体提取任务在社交媒体之中的应用，目的在于提升命名实体识别的准确率。在文本模态的基础之上又引入了图片模态，尝试利用图片特征补充文本上下信息，从而提高实体识别的准确率，并在此基础上形成了多模态实体识别的研究思路，且构建了与该任务相对应的多模态数据集。在这之后，Yu J. F. 等 [7] 提出基于 Transformer 架构的多模态命名实体识别模型；Zhang D.[8] 和 Zheng C. M.[9] 等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。Sun L. 等 [10] 通过多模态预训练模型和关系传播机制来实现命名实体识别任务并取得了很好的效果。Xu B. 等 [11] 提出一个多模态命名实体识别的通用框架，分为文本编码器、图像编码器、跨模态特征对齐单元、跨模态特征交互单元、跨模态特征匹配单元、跨模态特征融合单元、解码器，其中 BERT（bidirectional encoder representations from Transformers）[12] 用于文本编码，resNet[13] 用于图像编码，特征融合参考了 ViLBERT（vision and language BERT）[14] 模型，在效果上取得了较好提升。

诸如上述多模态命名实体识别任务已经有很多出色的工作，但模型结构大都过于复杂，需要做大量的前期准备工作。且目前，如何促进多种模态之间的融合仍是一个巨大挑战，随着跨模态注意力机制的提出以及各种多模态预训练模型的涌现，使得多模态数据之间的融合变得更为容易，本模型的创新性体现在：1）引入 CLIP（contrastire language-image pre-training）模型，相比传统的 ResNet 网络，CLIP 模型拥有更强的跨模态学习能力，可同时处理文本与视觉数据，学习文本与视觉信息之间的语义关系；2）加入跨模态的交叉注意力机制融合层，相比对多种模态的特征进行线性拼接具有更好的联合表示能力。

预训练模型 [15] 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间的语义对应关系，提升了视觉语言模型的泛化性和鲁棒性。预训练模型逐步扩展到多模态领域，相对于纯文本的预训练模型，多模态预训练模型 [15] 可以更好地对细粒度多模态语义单元间进行建模，从而保证模态之间的强关联性，这为将命名实体识别任务拓展到多模态场景提供了新思路。多模态预训练一般遵循同一个研究框架，以图像－文本为例，通常包括图像编码模块、文本编码模块、多模态特征融合模块、解码模块和预训练任务 5 个模块。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。CLIP[16] 模型是 Open AI 在 2021 年初发布的一种多模态预训练神经网络模型，被用于匹配图像和文本。该模型的关键创新之一是将图像和文本映射到统一的向量空间，通过对比学习的方式进行预训练，使模型能够直接在向量空间中计算图像和文本之间的相似性，无需额外的中间表示。

本文提出的基于视觉与语义增强的多模态命名实体识别模型，首先 CLIP 模型对数据集中的图文对进行预训练，然后使用 BERT 作为文本编码器获取输入文本的文本特征，用 CLIP 模型中的图像编码器作

为视觉编码器来获取输入图像的视觉特征，之后经由基于跨模态协同交叉注意力机制的特征融合单元对获取的文本特征和视觉特征进行融合，得到联合特征表示，最后将联合特征送入 CRF[17]（conditional random field）解码层得到整个句子的最优标注序列。

2 基于视觉与文本语义增强的多模态命名实体识别模型

本文提出的模型由文本编码、视觉编码、跨模态特征融合、CRF 解码等 4 个单元组成。文本编码单元由 BERT 编码器构成，对输入文本进行词嵌入，最后输出相应的文本特征。视觉编码单元由 CLIP 构成，对输入的视觉信息进行处理，获得相应的视觉特征。跨模态特征融合单元由跨模态交叉注意力机制构成，对生成的文本特征和视觉特征进行交叉注意力运算后得到交互特征，再对交互后的文本特征和视觉特征进行点积运算，得到跨模态的联合特征表示。最后，将得到的跨模态联合特征送入 CRF 解码层，以得到最初输入文本所对应的最优标注序列。

模型的整体结构如图 1 所示图中 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 为视觉到文本方向的查询（Query）、键（Key）和值（Value），上标 W、V 为向量维度。

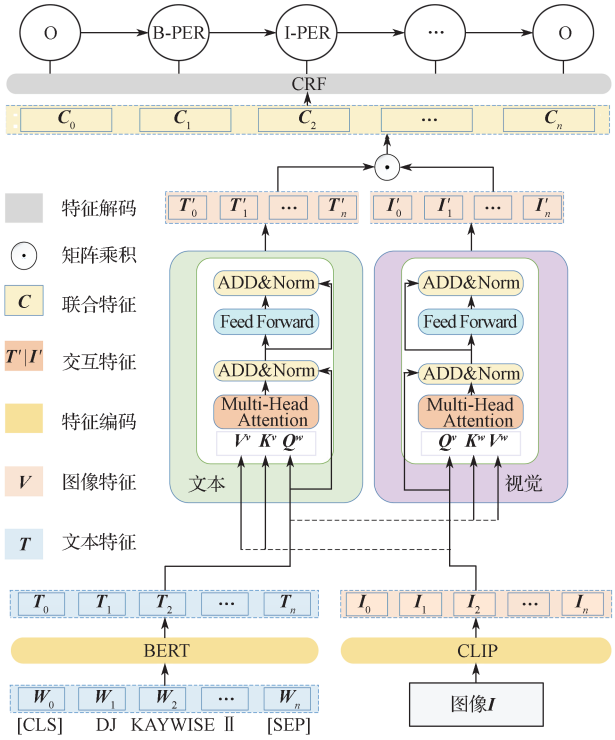

图 1 模型整体结构图

Fig. 1 Overall structure diagram of the model

2.1 特征编码

编码部分由 BERT、CLIP 组成。其中 BERT 用于编码数据集中的文本数据，CLIP 用于编码数据集中的图像数据。

2.1.1 文本编码

在自然语言处理领域，BERT 模型作为一种前沿的文本编码器，通过预训练获得的深层双向表示，为各种下游任务提供了强大的语言理解能力。BERT 的核心机制是基于 Transformer 架构的自注意力（self-attention）机制，能够有效地捕获文本中的长距离依赖关系。

对于数据集中的每条文本 $\boldsymbol{S}=[\boldsymbol{W}_0, \boldsymbol{W}_1, \cdots, \boldsymbol{W}_n]$，其中 $\boldsymbol{W}_i$ 为该条文本中的第 i 个单词，n 为文本的长度，BERT 首先对文本进行分词处理，文本会被 BERT 的分词器（tokenizer）分解成更细粒度的词元（tokens）。BERT 采用 WordPiece 分词算法，将未知或罕见单词分解为已知的子单元，以此提高模型的泛化能力。给定输入句子，其分词过程表示为

$$\boldsymbol{Input} = [\mathrm{CLS}] + tonkenizer(\boldsymbol{S}) + [\mathrm{SEP}]。 \quad (1)$$

式中：$tokenizer()$ 为 BERT 的分词函数；[CLS] 为句首的特殊标记符，经过 BERT 编码后该标记包含整条文本的语义信息；[SEP] 为句子结束的标记符。

分词时，对于文本中的每个单词 $\boldsymbol{W}_i$，若其不是最小文本单位，BERT 的分词器会将其进一步分解为最小单位的次元，例如，单词“lemon”将会被分为词元“le”和“mon”。

对于每个词元，BERT 将其转换为嵌入向量 $\boldsymbol{E}$（embeddings），这一过程包括 3 种嵌入的叠加：词元嵌入 $\boldsymbol{E}_{词元}$、位置嵌入 $\boldsymbol{E}_{位置}$和段落嵌入 $\boldsymbol{E}_{段落}$。

1）词元嵌入。将词元转换为固定长度的向量。

2）位置嵌入。由于 Transformer 不具备处理序列顺序的能力，位置嵌入确保了模型能够考虑到词元在句子中的位置信息。

3）段落嵌入（segment embeddings）。对于句子对输入，区分两个句子的嵌入。

综上所述，每个词元的最终嵌入表示为这 3 种嵌入的和：

$$\boldsymbol{E} = \boldsymbol{E}_{词元} + \boldsymbol{E}_{位置} + \boldsymbol{E}_{段落}。 \quad (2)$$

经过嵌入层处理后的向量输入到一系列 Transformer 编码层中。每个 Transformer 层主要包含两个子层：自注意力机制和前馈神经网络（feed forward neural network）。自注意力机制使模型能够在编码某个词元的表示时，考虑到句子中的其他词元。每个词元的嵌入向量通过 3 个不同的线性变换产生查询（$\boldsymbol{Q}$）、键（$\boldsymbol{K}$）和值（$\boldsymbol{V}$）向量。

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{Q}},\ \boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{K}},\ \boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{V}}。 \quad (3)$$

式中：$\boldsymbol{X}$ 为输入 $\boldsymbol{Input}$ 的嵌入向量；$\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$ 均为可学习的权重矩阵。

之后，模型计算每个词元（通过其查询向量）对

所有词元（通过其键向量）的注意力得分，这个得分决定了在编码当前词元时应该给予其他词元多少重视。计算公式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}。\quad（4）$$

式中$\sqrt{d_k}$为键向量的维度，这个归一化因子有助于防止点积的结果过大而导致 *softmax* 函数进入饱和区域，从而降低梯度的有效性。

通过这个过程，自注意力机制允许模型动态地调整每个词元的表示，使其不仅反映词元本身的信息，还包括了整个句子的上下文信息。在自注意力机制和前馈神经网络的每个子层后，BERT 采用残差连接和层归一化来提高训练的稳定性和效率。残差连接允许模型的输入直接加到子层的输出上，有助于缓解深层网络中的梯度消失问题。层归一化则是对每个子层输出的标准化处理，确保网络层输出的分布保持稳定。这可以通过以下公式表示：

$$LayerNorm\left(\boldsymbol{x}+Sublayer\left(\boldsymbol{x}\right)\right)。\quad（5）$$

式中：$\boldsymbol{x}$ 为子层的输入；*Sublayer*($\boldsymbol{x}$) 为子层自身的操作，比如自注意力或前馈神经网络的输出。

BERT 中的每个 Transformer 编码层中的前馈神经网络（FFN）可以对自注意力层的输出进行进一步处理。FFN 由两个线性变换组成，中间有一个 ReLU 激活函数：

$$FFN(\boldsymbol{x})=\max\left(0,\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1\right)\boldsymbol{W}_2+\boldsymbol{b}_2。\quad（6）$$

式中：$\boldsymbol{x}$ 为自注意力层的输出；$\boldsymbol{W}_1$、$\boldsymbol{W}_2$、$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 均为前馈网络参数。

对于数据集中的每一条文本，将其输入 BERT 进行编码并得到对应的嵌入向量，采用每个 ***Input*** 中 [CLS] 位置对应的向量作为最终的文本向量 $\boldsymbol{E}_T$，并将其模态融合层与视觉特征相融合。文本编码器结构图如图 2 所示。

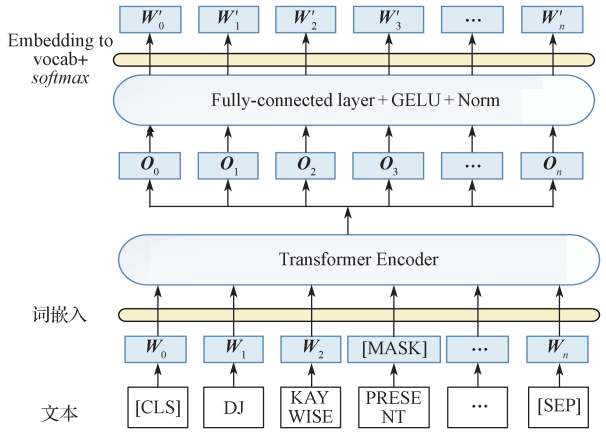

图 2　文本编码结构图

Fig. 2　Text encoding structure diagram

2.1.2　*视觉编码*

CLIP 模型通过预训练获得的跨模态表示，显著提高了机器对视觉内容的理解能力。CLIP 模型由两个主要部分组成：一个图像编码器和一个文本编码器。图像编码器负责将输入的图像转换为高维特征向量，而文本编码器则将文本信息（如标签或描述）转换为与视觉特征向量相同空间的向量。CLIP 通过对比学习框架进行预训练，优化模型以正确匹配图像和对应的文本描述。对比学习旨在最大化正样本（匹配的图像－文本对）的相似性，同时最小化负样本（不匹配的图像－文本对）的相似性。具体的损失函数可以表示为

$$L=-\log\left(\frac{\exp\left(sim\left(\boldsymbol{E}_I,\boldsymbol{E}_T\right)/\tau\right)}{\sum_{T\in T_{\text{batch}}}\exp\left(sim\left(\boldsymbol{E}_I,\boldsymbol{E}_T\right)/\tau\right)}\right)。\quad（7）$$

式中：$\boldsymbol{E}_I$ 为视觉特征向量；$\boldsymbol{E}_T$ 为与图像匹配的文本特征向量；$\boldsymbol{T}_{\text{batch}}$ 为一个包含正样本和负样本的文本集合；*sim*(•) 为计算相似性的函数；τ 为温度参数，用于调整相似性分数的尺度。

课题组利用 CLIP 通过跨模态预训练获得的强大的图像编码能力，将 CLIP 的图像编码器作为本文数据集中图像数据的编码器。

对于数据集中的每一个与文本数据相对应的图像数据，首先对其进行预处理以符合图像编码器的输入格式。预处理步骤通常包括图像尺寸调整以及像素归一化，设 $\boldsymbol{I}$ 为原始输入图像，预处理后的图像表示为 $\boldsymbol{I}'$，接下来，预处理后的图像 $\boldsymbol{I}'$ 被分割成一系列的小块（patches），每个块被映射到高维空间，作为 Transformer 的输入：

1）分块。图像 $\boldsymbol{I}'$ 被分割成 N 个大小相等的块，设图像分块表示为 $\{\boldsymbol{P}_0,\boldsymbol{P}_1,\cdots,\boldsymbol{P}_N\}$。

2）块嵌入。每个图像块 $\boldsymbol{P}_i$ 通过一个线性投影层转换为一个固定维度的嵌入向量 $\boldsymbol{E}_i$。此外，为每个嵌入向量加上位置嵌入 $\boldsymbol{Pos}_i$ 以保留空间信息：

$$\boldsymbol{E}_i=\boldsymbol{P}_i\boldsymbol{W}+\boldsymbol{Pos}_i。\quad（8）$$

式中：$\boldsymbol{W}$ 为线性投影的权重矩阵；$\boldsymbol{Pos}_i$ 为该块对应的位置嵌入。

经过块嵌入后，可以得到每个图像的嵌入向量：

$$\boldsymbol{E}=\{\boldsymbol{E}_0,\boldsymbol{E}_1,\cdots,\boldsymbol{E}_N\}。\quad（9）$$

在得到特征向量以后，需要将其输入到 CLIP 的 Transformer 层进行编码，Transformer 的结构与 2.1.1 章节中所介绍的相同。在经过多层 Transformer 处理后，通常采用“类别”（class）token 的最终输出作为图像的整体特征表示。假设 $\boldsymbol{C}$ 是最后一个 Transformer 层中类别 token 的输出向量，$\boldsymbol{C}$ 即作为图

像 $\boldsymbol{I}$ 的向量表示，即：

$$\boldsymbol{E}_{\boldsymbol{I}}=\boldsymbol{C}。$$

最终，图像 $\boldsymbol{I}$ 通过 CLIP 的视觉编码被编码为一个高维向量 $\boldsymbol{E}_{\boldsymbol{I}}$。

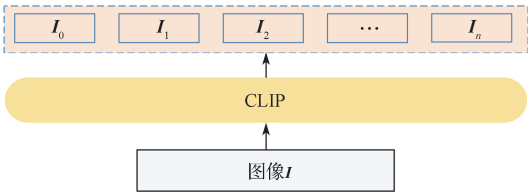

图 3 视觉编码结构图

Fig. 3 Visual encoding structure diagram

2.2 基于协同交叉注意力机制的跨模态特征融合

通过文本编码器和视觉编码器获得文本和图像对应的嵌入向量之后，采用协同交叉注意力机制（co-attention mechanism）实现跨模态数据之间的信息交互与共享。

设经过编码器编码后的图像和文本特征（嵌入向量）分别为

$$\boldsymbol{E}_{\boldsymbol{T}}=\{\boldsymbol{E}_{t0},\boldsymbol{E}_{t1},\cdots,\boldsymbol{E}_{tn}\},\tag{10}$$

$$\boldsymbol{E}_{\boldsymbol{I}}=\{\boldsymbol{E}_{i0},\boldsymbol{E}_{i1},\cdots,\boldsymbol{E}_{im}\}。\tag{11}$$

协同交叉注意力机制将 $\boldsymbol{E}_{\boldsymbol{T}}$ 和 $\boldsymbol{E}_{\boldsymbol{I}}$ 作为输入，对于从图像到文本的交互，设定视觉特征 $\boldsymbol{E}_{ij}$ 为 Query，文本特征 $\boldsymbol{E}_{ti}$ 同时作为 Key 和 Value，跨模态融合后的文本特征表示为

$$\boldsymbol{E}_{ti}^{'}=CoAttention_{\boldsymbol{I}\to\boldsymbol{T}}\left(\boldsymbol{E}_{ij},\boldsymbol{E}_{ti}\right)。\tag{12}$$

对于从文本到图像的交互，设定文本特征 $\boldsymbol{E}_{ti}$ 为 Query，视觉特征 $\boldsymbol{E}_{ij}$ 同时作为 Key 和 Value，跨模态融合后的文本特征表示为

$$\boldsymbol{E}_{ij}^{'}=CoAttention_{\boldsymbol{T}\to\boldsymbol{I}}\left(\boldsymbol{E}_{ti},\boldsymbol{E}_{ij}\right)。\tag{13}$$

具体来说，跨模态协同交叉注意力机制的计算过程首先要计算跨模态注意力分数，对于图像到文本的方向，注意力分数 $Score_{\boldsymbol{I}\to\boldsymbol{T}}$ 的计算式为

$$Score_{\boldsymbol{I}\to\boldsymbol{T}}=softmax\left(\frac{\left(\boldsymbol{E}_{ij}\boldsymbol{W}_{\boldsymbol{I}\to\boldsymbol{T}}^{\boldsymbol{Q}}\right)\left(\boldsymbol{E}_{ti}\boldsymbol{W}_{\boldsymbol{I}\to\boldsymbol{T}}^{\boldsymbol{K}}\right)^{\mathrm{T}}}{\sqrt{d_k}}\right)。\tag{14}$$

在计算得到注意力分数之后，需要利用注意力分数对特征进行加权，得到加权的文本特征表示 $\boldsymbol{E}_{ti}^{'}$ 为

$$\boldsymbol{E}_{ti}^{'}=Score_{\boldsymbol{I}\to\boldsymbol{T}}\cdot\left(\boldsymbol{E}_{ti}\boldsymbol{W}_{\boldsymbol{I}\to\boldsymbol{T}}^{\boldsymbol{V}}\right)。\tag{15}$$

式（14）（15）：$\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$ 和 $\boldsymbol{W}^{\boldsymbol{V}}$ 分别为图像到文本方向的查询（Query）、键（Key）和值（Value）的变换矩阵；$\sqrt{d_k}$ 为键向量的维度，用于点积结果的缩放，以稳定梯度。

将 $\boldsymbol{E}_c=\boldsymbol{E}_t^{'}$ 作为数据集中的图像和文本数据融合后最终得到的跨模态向量，作为解码器的输入。

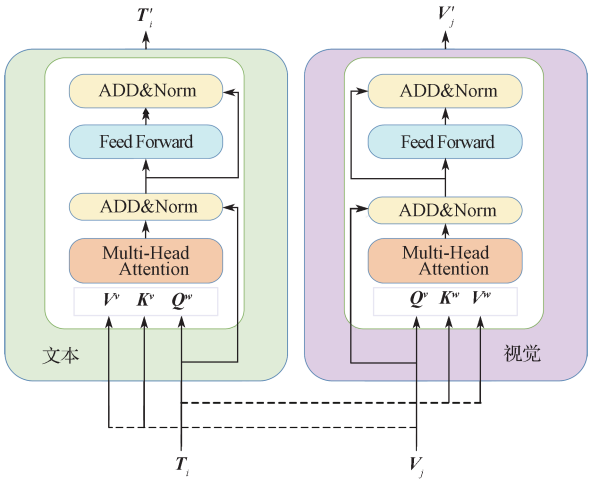

图 4 跨模态特征融合结构图

Fig. 4 Cross-modal feature fusion structure diagram

2.3 特征解码

本文采用条件随机场对文本的跨模态向量进行解码。CRF 是一种用于预测序列数据标签的统计建模方法，能够在整个序列层面上考虑标签之间的依赖关系，为整个序列寻找最优的标签配置，而不是独立地为每个位置选择最佳标签，因此 CRF 在处理命名实体识别任务时具有显著的优势。

在 CRF 层中，每个序列位置 $\boldsymbol{I}$ 的标签 y_i 不仅取决于该位置的输入特征，还取决于相邻位置的标签。这种依赖关系通过转移矩阵 $\boldsymbol{A}$ 来建模，其中矩阵的元素 $A_{k,l}$ 表示从标签 k 转移到标签 l 的转移得分。

给定跨模态向量表示 $\boldsymbol{E}_c=\{\boldsymbol{E}_{c0},\boldsymbol{E}_{c1},\cdots,\boldsymbol{E}_{cn}\}$，其中 $\boldsymbol{E}_{ci}$ 表示该序列中的第 i 个输入特征。整个序列的得分函数定义为

$$S(\boldsymbol{E}_c,\ y)=\sum_{i=1}^{n}\left(\boldsymbol{W}_{yi}\cdot\boldsymbol{E}_{ci}+\boldsymbol{A}_{y_{i-1},\,y_i}\right)。\tag{16}$$

式中：$\boldsymbol{W}_{yi}$ 为与标签 y_i 相关的特征权重向量；$\boldsymbol{A}_{y(i-1),\,y(i)}$ 为从标签 y_{i-1} 转移到标签 y_i 的转移得分。

为了将得分转换为概率，需要对所有可能的标签序列 $y^{'}$ 上的得分进行归一化：

$$P(y\mid\boldsymbol{E}_c)=\frac{\exp\left(S(\boldsymbol{E}_c,\ y)\right)}{Z(\boldsymbol{E}_c)}。\tag{17}$$

式中，$Z(\boldsymbol{E}_c)$ 为归一化因子，通过对所有可能的标签序列 $y^{'}$ 的得分求和并取指数得到：

$$Z(\boldsymbol{E}_c)=\sum_{y^{'}}\exp\left(S\left(\boldsymbol{E}_c,y^{'}\right)\right)。\tag{18}$$

之后，解码过程的目标是找到给定输入 $\boldsymbol{E}_c$ 下最有可能的序列标签 y^{*}，通过维特比算法来实现，该算法通过动态规划来有效地搜索最优路径，即最有可能的标签序列：

$$y^{*}=argmax_{y}P\left(y\mid \boldsymbol{E}_{t}\right)。\quad (19)$$

通过上述过程，CRF 层能够有效地考虑标签之间的转移概率，为跨模态向量表示的文本序列提供精确的标签预测，从而在本文的序列标注任务中实现高性能。

3　实验设置

3.1　实验数据集

为了证明本文所提方法的有效性，本文采用 TWITTER-2015、TWITTER-2017 两个多模态数据集作为实验数据集，数据集内容如表 1 所示。

表 1　两种多模态数据集内容统计

Table 1　Statistics of two multi-modal datasets content

类别	TWITTER-2015			TWITTER-2017		
	Train	Dev	Test	Train	Dev	Test
PER	2 217	552	1 816	2 943	626	621
LOC	2 091	522	1 697	731	173	178
ORG	928	247	839	1 674	375	395
MISC	940	225	726	701	150	157
Total	6 176	7 546	5 075	6 049	1 324	1 351

数据集分为 PER（人名）、LOC（地点）、ORG（组织或机构）、MISC（其他实体）等 4 种实体类型。

3.2　实验环境

本实验基于 PyTorch 框架搭建模型，并使用 GPU 计算框架进行加速训练。详细的硬件配置如表 2 所示。

表 2　硬件配置表

Table 2　Hardware configuration table

系统配置	机器环境
操作系统	Windows 11 × 64 位
CPU	12th Core i7-12700H
GPU	NVIDIA GeForce RTX 3060
内存	16 GB
Python 版本	3.9.17
PyTorch 版本	1.12.1

3.3　基线模型

为了验证本模型，特设计如下对比实验方案：

1）BiLSTM（bidirectional LongShort-Term memory）+CRF 模型。由双向长短期记忆网络 BiLSTM 和条件随机场组合而成。BiLSTM 是一种适合处理序列数据的循环神经网络（RNN）变体，具有双向的记忆能力，能够捕捉输入序列中的上下文信息。CRF 是一种用于序列标注的概率图模型，能够显式地建立模标签之间的依赖关系，有利于学习序列标签中的全局结构。在此模型中，BiLSTM 用于提取输入句子的特征表示，并将这些特征表示作为输入传递给 CRF 层，用于对每个词进行标签预测。整个模型在训练过程中会同时学习特征提取和标签预测这两个任务，以最大化标签序列的联合概率。

2）BERT-CRF 模型。使用 BERT-CRF 模型进行命名实体识别时，首先，将文本输入 BERT 模型中，获取句子的表示向量；然后，将得到的表示向量输入 CRF 模型中，进行标签预测，最终获得命名实体的识别结果。

3）GVATT-BERT-CRF 模型。该模型是一种结合了 Global Attention Mechanism 和 BERT 预训练模型以及条件随机场（CRF）的命名实体识别模型，引入了全局自注意力机制，能够更好地捕捉文本中的全局语义信息。在 GVATT-BERT-CRF 模型中，首先使用 BERT 模型对输入文本进行特征提取，得到文本的表示向量。随后，利用 Global Attention Mechanism 方法将 BERT 输出的表示向量进行进一步的特征集成，以更好地捕捉句子的全局语义信息。最后，将集成后的特征输入 CRF 模型中进行标签预测，从而完成命名实体识别任务。GVATT-BERT-CRF 模型在命名实体识别任务中能够取得较好的性能表现，特别是在处理长文本、多实体和上下文相关性较强的情况下具有优势。

4）AdaCAN-BERT-CRF 模型。AdaCAN 是一种自适应上下文感知注意力网络，能够有效地捕捉文本中的上下文信息，有助于提升命名实体识别的准确性和鲁棒性。因此，AdaCAN-BERT-CRF 模型可以充分利用自适应上下文感知关注机制、BERT 的语言表示学习能力和 CRF 的序列标注优势，实现高效的命名实体识别。

5）本模型。采用 BERT 作为文本编码，CLIP 中的图像编码器作为视觉编码器，使用跨模态协同交叉注意力机制对由经过文本编码和图像编码输出的特征进行融合，从而得到多模态特征的联合表示，再将联合特征送入 CRF 层，最后得到预测序列。

4　实验结果

4.1　超参数设置

本实验将最大序列长度设置为 40，batch_size 为 32，文本编码单元采用 bert-based-cased 模型，视觉编码单元采用 openai/clip-vit-base-patch 模型，其他参数详见表 3。

表 3 超参数设置表

Table 3 Hyper-parameter setting

参数	取值	参数	取值
num_epochs	40	*eval_begin_epoch*	1
batch_size	32	*max_seq*	40
lr	5e−5	*hidden_dropout_prob*	0.1
warmup_ratio	0.06		

其中共设置 *num_epochs*（epoch 数量）、*batch_size*（批大小）、*lr*（学习率）、*warmup_ratio*（预热初始化学习率）、*eval_begin_epoch*（起始 epoch）、*max_seq*（最大序列长度）、*hidden_dropout_prob*（隐藏层 dropout 概率）7 个参数。

4.2 模型对比实验

本文采用准确率、召回率、F_1 值作为模型的评判标准。分别在 TWITTER-2015 和 TWITTER-2017 两个数据集上进行实验，并采用准确率、召回率和 F_1 值等作为实验结果评判标准，对比实验结果如表 4、表 5 所示。

表 4 在 TWITTER-2015 上的实验结果对比

Table 4 Comparison of experimental results of TWITTER-2015 %

模型	*P*	*R*	F_1
BiLSTM-CRF	68.14	61.09	64.42
BERT-CRF	69.22	72.37	70.76
GVATT-BERT-CRF	69.15	72.59	70.83
AdaCAN-BERT-CRF	69.87	73.45	70.62
本模型	70.15	73.81	71.93

表 5 在 TWITTER-2017 上的实验结果对比

Table 5 Comparison of experimental results of TWITTER-2017 %

模型	*P*	*R*	F_1
BiLSTM-CRF	79.42	73.43	76.31
BERT-CRF	83.32	83.57	83.44
GVATT-BERT-CRF	83.64	84.38	84.01
AdaCAN-BERT-CRF	85.13	83.20	84.15
本模型	83.97	84.46	84.21

由表 4 可知，本文提出的模型，较 GVATT-BERT-CRF 在 4 种类别的实体进行的实验中 F_1 值提升了 1.10%，召回率提升 1.22%。较 AdaCAN-BERT-CRF 在实验中 F_1 值提升 1.31%，召回率提升 0.36%。

由表 5 可知，本文提出的模型，较 GVATT-BERT-CRF 在 4 种类别的实体进行的实验中 F_1 值提升了 0.20%，召回率提升 0.08%。较 AdaCAN-BERT-CRF 在实验中 F_1 值提升 0.06%，召回率提升 1.26%。

综上所述，和以上各个模型相比，本文方法有显著提升。由实验分析可知，本文中的方法在视觉和文本模态的融合任务上取得了一定的效果，促进了视觉信息与文本信息之间的语义补充。

5 结论

为了使视觉信息能够更好地对文本语义信息进行补充，本文提出了基于视觉文本语义增强的多模态命名实体识别方法，它分别对输入的文本和图像进行处理，从而获得相应的特征表示，再将文本特征和视觉特征经由基于跨模态协同交叉注意力机制的特征交互单元，以得到跨模态的联合特征表示，本方法结合了预训练模型的图像表示能力与跨模态协同交叉注意力机制的特征交互能力，相比传统方法取得了显著提升。

后续考虑从以下方面优化本方法：

1）对数据集进行增强处理，设计数据增强单元，采用目标检测领域中的视觉检测技术对原始数据集进行处理，以得到更好的效果。

2）增加特征匹配单元，在将文本特征和图片特征送入融合单元之前对其进行一次特征对齐。

3）引入更为先进的多模态预训练模型，用于特征增强。

参考文献：

[1] 陈 烨，周 刚，卢记仓 . 多模态知识图谱构建与应用研究综述 [J]. 计算机应用研究，2021，38(12)：3535−3543.
CHEN Ye，ZHOU Gang，LU Jicang. Survey on Construction and Application Research for Multi-Modal Knowledge Graphs[J]. Application Research of Computers，2021，38(12)：3535−3543.

[2] 孙影影，贾振堂，朱昊宇 . 多模态深度学习综述 [J]. 计算机工程与应用，2020，56(21)：1−10.
SUN Yingying，JIA Zhentang，ZHU Haoyu. Survey of Multimodal Deep Learning[J]. Computer Engineering and Applications，2020，56(21)：1−10.

[3] 韩 普，陈文祺 . 多模态命名实体识别研究进展 [J]. 数据分析与知识发现，2024，8(4)：50−63.
HAN Pu，CHEN Wenqi. Review of Multimodal Named Entity Recognition Studies[J]. Data Analysis and Knowledge Discovery，2024，8(4)：50−63.

[4] FENSEL D，ŞIMŞEK U，ANGELE K，et al. Introduction：What Is a Knowledge Graph? [M]// Knowledge Graphs. Cham：Springer International Publishing，2020：1−10.

[5] WANG M，QI G L，WANG H F，et al. Richpedia：

A Comprehensive Multi-Modal Knowledge Graph[M]// Lecture Notes in Computer Science. Cham：Springer International Publishing，2020：130−145.

[6] MOON S，NEVES L，CARVALHO V. Multimodal Named Entity Recognition for Short Social Media Posts[J]. ArXiv e-Prints，2018：1802.07862.

[7] 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. Stroudsburg：Association for Computational Linguistics，2020：3342−3352.

[8] ZHANG D，WEI S Z，LI S S，et al. Multi-Modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2021，35(16)：14347−14355.

[9] 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，2021，23：2520−2532.

[10] SUN L，WANG J Q，ZHANG K，et al. RpBERT：A Text-Image Relation Propagation-Based BERT Model for Multimodal NER[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2021，35(15)：13860−13868.

[11] XU B，HUANG S Z，SHA C F，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. [S. l.]：ACM，2022：1215−1223.

[12] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-Training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. [2024−03−01]. https://arxiv.org/abs/1810.04805v2.

[13] HE K M，ZHANG X Y，REN S Q，et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas：IEEE，2016：770−778.

[14] LU J，BATRA D，PARIKH D，et al. ViLBERT：Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver：Curran Associates，2019：13−23.

[15] 李　耕，王梓烁，何相腾，等 . 从 ChatGPT 到多模态大模型：现状与未来 [J]. 中国科学基金，2023，37(5)：723−724.
LI Geng，WANG Zishuo，HE Xiangteng，et al. From ChatGPT to Multimodal Large Models：Current Situation and Future[J]. Bulletin of National Natural Science Foundation of China，2023，37(5)：723−724.

[16] RADFORD A，KIM J W，HALLACY C，et al. Learning Transferable Visual Models from Natural Language Supervision[J]. ArXiv e-Prints，2021：2103.00020.

[17] HUANG Z H，XU W，YU K. Bidirectional LSTM-CRF Models for Sequence Tagging[EB/OL]. [2024−03−01]. https://arxiv.org/abs/1508.01991v1.

（责任编辑：申　剑）