计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100071

基于量子 Transformer 的多模态实体关系联合抽取方法

李代祎　孔德龙　吴怀广　张佳慧　韩宇璨
郑州轻工业大学计算机科学与技术学院　郑州 450000
(lidaiyiyy@163.com)

摘　要　多模态命名实体识别(Multimodal Name Entity Recognition,MNER)和多模态关系抽取(Multimodal Relation Extraction,MRE)是多模态知识图谱构建中的两个关键技术。然而,现有的 MNER 和 MRE 方法在对高维数据进行特征提取和融合时还存在一定的局限性。为了解决这些问题,提出了一种基于量子 Transformer 的多模态实体关系联合抽取方法。首先,设计一种针对文本数据处理的参数化量子电路,该线路利用量子力学中的叠加和纠缠特性,结合 Transformer 模型提取文本深层特征;其次,通过设计的金字塔视觉特征提取模型获取包含从高到底的金字塔状的层次特征,充分考虑到了图像的多尺度信息。最后,通过设计的分层视觉前缀网络将分层多尺度图像特征与文本特征对齐并融合,获取鲁棒性高的文本表示。本研究为多模态实体关系抽取提供了新的研究思路,在 3 个公开基准数据集上的实验结果表明,提出的基于量子 Transformer 多模态实体关系抽取方法是有效且稳定的。

关键词:多模态实体识别;多模态关系抽取;金字塔特征;Transformer;特征融合
中图分类号　TP391

Multimodal Entity-Relation Joint Extraction Method Based on Quantum Transformer

LI Daiyi,KONG Delong,WU Huaiguang,ZHANG Jiahui and HAN Yucan
College of Computer Science and Technology,Zhengzhou University of Light Industry,Zhengzhou 450000,China

Abstract　Multimodal Name Entity Recognition(MNER) and Multimodal Relation Extraction(MRE) are two key technologies in the construction of multimodal knowledge graphs. However,the existing MNER and MRE methods still have certain limitations in feature extraction and fusion of high-dimensional data. To address these issues,this paper proposes a multimodal entity relation joint extraction method based on quantum Transformer. Firstly,a parameterized quantum circuit for text data processing is design,which utilizes the superposition and entanglement characteristics in quantum mechanics,and combines with the Transformer model to extract deep features from text;Secondly,the pyramid visual feature extraction model is designed to obtain hierarchical features from high to low,which fully considers the multi-scale information of the image. Finally,by designing a hierarchical visual prefix network,the hierarchical multi-scale image features are aligned and fused with the text features to obtain a highly robust text representation. This study provides a new research approach for multimodal entity relation joint extraction. Experimental results on three public benchmark datasets show that the multimodal entity relation extraction method based on quantum Transformer proposed in this paper is effective and stable.

Keywords　MNER,MRE,Pyramid visual feature,Transformer,Feature fusion

1　引言

多模态实体和关系抽取是自然语言处理领域中的两个关键子任务,旨在通过关联图像来增强文本数据的语义表征从而进行消歧,进而提高多模态命名实体识别(MNER)模型和关系抽取(MRE)模型性能。与传统单模态实体识别和关系抽取方法[1,2]不同,MNER 和 MRE 任务主要面临两方面的挑战:1)如何有效获取与文本中实体相关的视觉信息;2)如何减少图像中冗余信息对 NER 模型的影响。如图 1 所示,文本中的“Oscars”可能是一个人名或是一个电影奖。然而,图像中奖杯信息可以补充文本语义的不足,从而确定“Oscars”是一个电影奖。此外,图像中已知“持有”是人和奖杯之间的关系,那么就可以理解实体“Attenborough”和实体“Oscars”之间的关系为“awarded”。因此,针对 MNER 和 MRE 任务,从图像中捕获与实体相关的视觉信息是必不可少的且具有挑战性。

目前,一些优秀的 MNER 和 MRE 方法被提出。例如,Moon 等[3]提出借助图像的视觉特征来增强文本的语义表征,有效提高了基于文本的信息抽取模型;Zheng 等[4]进一步

基金项目:国家自然科学基金(61672470);河南省重大科技专项(超导量子芯片设计与制备关键技术研究)项目(221100210400);河南省重大公益项目(201300210200);郑州轻工业大学博士基金项目(2024BSJJ014)
This work was supported by the National Natural Science Foundation of China(61672470),Major Science and Technology Research Projects in Henan Province(221100210400),Major Public Welfare Projects in Henan Province,China(201300210200) and Doctoral Research Fund of Zhengzhou University of Light Industry(2024BSJJ014).
通信作者:吴怀广(hgawu@126.com)

验证了对象视觉特征与文本融合对于 MNER 和 MRE 更具特异性和重要性；Sun 等[5]提出在 MNER 任务之前训练一个分类器来判断“图像是否增加了文本的含义”，有效提高了模型性能。然而，该方法严重依赖于对大量额外标注的图像-文本相关性语料的预训练，只关注整个图像，而忽略了不相关视觉特征的误差影响。事实上，不相关视觉特征可能会直接对文本推理产生负面影响。因此，在 MNER 和 MRE 任务中，需要一种有效的方法来学习更好的视觉表示，以减少图片中不相关视觉特征的误差影响。

Text: Attenborough and Ben Kingsley with their Oscars.

实体1	类型	实体2	类型	关系
Attenborough	Per	Oscars	Misc	awarded
Ben Kingsley	Per	Oscars	Misc	awarded
Attenborough	Per	Ben Kingsley	Per	peer

图 1 文本和图像对的实例分析

Fig. 1 Example analysis of text and image pairs

为了捕获有效的视觉特征，受到提示学习的启发[6-8]，一些研究者将图像中的对象视觉特征前置到 BERT 模型的 Transformer 层的文本序列前。这个视觉前缀是一个可插拔的操作，不需要对相关性进行注释。另外，一些研究表明，卷积神经网络(Convolutional Neural Networks，CNN)模型可以从图像中获取包含从低到高的金字塔状的多尺度特征信息。例如，Han 等[9]认为不同大小的图像可以在相应尺度上有合适的特征表示。受此启发，本文提出在 BERT 的 Transformer 层感知分层的多尺度视觉特征，从而使得模型做出更准确的预测。

此外，随着量子技术的发展，量子力学结合深度学习的方法被引入到多模态数据分析中，为建模不同模态数据之间的复杂交互提供了一种新方法。例如，Tiwari 等[10]在模态内部以及跨模态之间进行交互建模，解决了传统方法难以有效捕捉句子中内在语义差异和复杂关联问题；Phukan 等[11]提出了一个基于量子启发的多模态情感分析(QMSA)框架，该框架利用量子力学原理对模态之间的关联进行建模，有效捕获到了更复杂和高层次的多模态交互信息。然而，尽管基于量子的深度学习方法在多模态数据特征提取上效果显著，但该方法在多模态实体关系抽取领域仍面临诸多挑战。例如，量子启发方法与经典机器学习技术的有效集成、处理数据噪声，以及提高模型的可解释性和透明性等。

为了解决上述问题，本文提出了一种基于量子 Transformer 的多模态实体关系联合抽取方法。首先，设计了一种面向文本的参数化量子线路，该线路利用量子叠加和量子纠缠特性，实现对文本数据的深层特征提取。其次，设计视觉前缀引导的融合机制，将相关的视觉特征作为 Transformer 层的文本前缀，这使得视觉增强的实体识别和关系抽取模型的鲁棒性更强。最后，采用 CNN 抽取图像特征，获取包含从高到底的金字塔状的层次特征，并将各种聚合的分层多尺度视觉特征作为增强实体识别和关系抽取的视觉前缀。本文的主要贡献如下：

1)提出了一种面向文本的参数化量子线路，并将该线路与多头自注意力机制结合，用于对文本特征进行深层次的特征提取。

2)设计了一个分层视觉前缀融合网络，在 BERT 的 Transformer 层通过基于视觉前缀的注意机制将分层多尺度视觉特征与文本特征融合在一起，以生成有效和鲁棒性强的文本语义表示。

3)构建了一个金字塔视觉特征提取模型捕获包含从高到底的金字塔状的层次特征，使得 Transformer 层中的文本表示都可以有效感知相应的分层视觉特征。

2 相关工作

多模态实体和关系抽取的早期研究侧重于单一模态数据的信息抽取[12-13]。近年来，随着多模态数据在社交媒体平台上急速增加，一些研究集中在 MNER 和 MRE 任务上，旨在通过关联图像更好地识别文本中的命名实体及实体间关系。

在早期阶段，一些研究者[14-15]提出采用 RNN 对文本进行编码，并通过 CNN 对整个图像进行编码，然后设计隐式交互方式来对两种模态之间的信息进行建模，以探索多模态 NER 任务。随着多模态数据处理技术的发展，Yu 等[16]提出利用区域图像特征来表示图像中的对象，与基于 Transformer 的文本特征进行语义融合，有效提高了多模态实体识别模型的性能。

目前，尽管一些研究者提出通过学习文本-图像关系分类器来增强多模态 BERT 模型[5]，以减少不相关图像的干扰，但是这些方法需要对图像-文本对的不相关性进行大量注释。现有的视觉-语言 BERT 模型主要体现在模型架构和预训练任务上。1)模型架构不同。单流结构有 unicode-vl[17]，VisualBERT[18]，VL-BERT[19]和 UNITER[20]等，其中文本和图像被组合成一个序列并输入 BERT 以学习上下文嵌入表示。双流结构有 LXMERT[21]和 ViLBERT[22]，它们分别将视觉和语言处理成两个流，并通过跨模态或共注意力转换层进行融合。2)预训练任务不同。多模态视觉语言模型的预训练任务主要包括掩模语言建模(Masked Language model，MLM)、掩模区域分类(Masked Region Classification，MRC)和图像-文本匹配(Image-Text Matching，ITM)。然而，当前的模型主要侧重于理解图像和文本之间的对应关系，而 MNER 和 MRE 任务更强调利用视觉信息来增强文本的理解。因此，将当前的视觉语言模型应用于 MNER 和 MRE 任务上仍需进一步优化和改进。

3 基于量子 Transformer 的多模态实体关系抽取

3.1 多模态实体关系抽取框架图

本文提出了一个基于量子 Transformer 的多模态实体关系联合抽取方法，如图 2 所示，主要包括基于量子 Transformer

的文本处理模块、面向图像的金字塔特征提取模块和基于视觉前缀网络的图文特征融合模块。

首先，基于提出的量子 Transformer 模型实现对文本数据的深层特征提取。其次，通过设计的金字塔视觉特征提取模型获取包含从高到底的金字塔状的层次特征，充分考虑到了图像的多尺度信息。然后，在 BERT 的 Transformer 计算框架中，通过设计的分层视觉前缀网络将分层多尺度特征与文本特征融合，获取有效且鲁棒性强的文本表征。最后，基于视觉特征感知的文本表征，采用 CRF 获取文本中的实体关系标注序列。

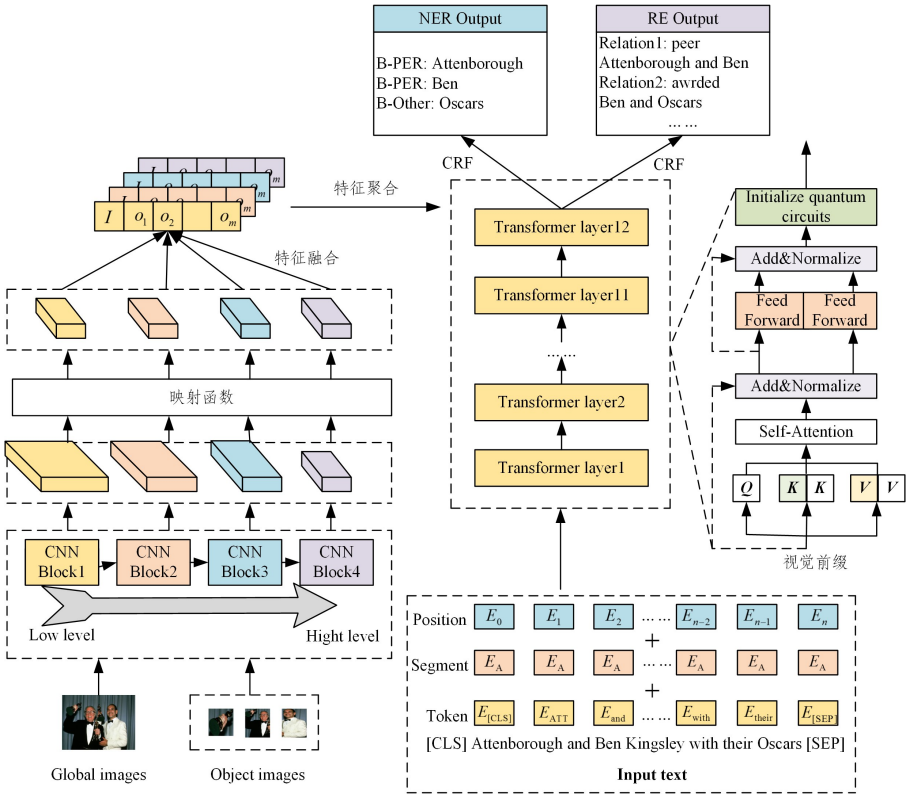


图 2　基于量子 Transformer 的多模态实体关系联合抽取框架

Fig. 2　Multimodal entity-relationship joint extraction framework based on quantum Transformer

3.2　基于量子 Transformer 的文本特征提取

量子计算以其在解决复杂计算问题上的潜力而备受关注。本文构建的量子文本学习模块是多模态实体关系抽取模型中的核心组成部分，主要负责提取文本数据中实体和关系的相关特征。该模块基于 BERT 预训练模型，通过与量子 Transformer 的结合，增强模型对文本的理解深度和特征表示能力。具体如图 3 所示。

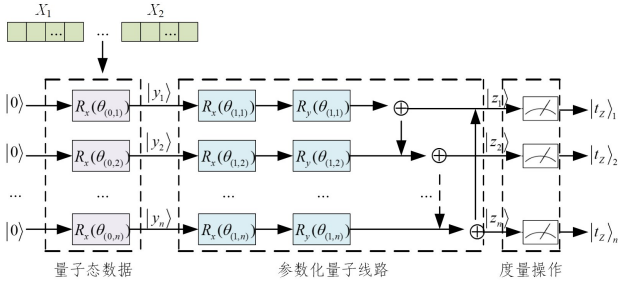


图 3　面向文本的参数化量子电路

Fig. 3　Text oriented parameterized quantum circuit

BERT 模型作为一个高效的语言处理预训练模型，被用于原始文本的特征提取。假设原始文本表示为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，其中 x_i 表示输入序列中第 i 个 Token。通过 BERT 嵌入层获取包含深层语义信息的 Token 嵌入序列 $\boldsymbol{H}^l=\{h_1^l,h_2^l,\cdots,h_n^l\}$，$\boldsymbol{H}^l\in\mathbb{R}^{n\times d}$，其中 h_i^l 表示输入序列在第 l 层 Transformer 中第 i 个 Token 的嵌入表示，n 表示输入序列长度，d 表示隐藏层维数，l 表示 Transformer 的层数，则第 $L-1$ 层的输出如下所示：

$$\boldsymbol{H}^{L-1}=\{h_1^{L-1},h_2^{L-1},\cdots,h_n^{L-1}\},\boldsymbol{H}^{L-1}\in\mathbb{R}^{n\times d} \tag{1}$$

量子 Transformer 的设计是结合量子计算的特性来增强模型深层特征提取的能力。如图 3 所示，通过量子计算技术来引导传统 Transformer 中的多头注意力机制建模。具体地，将输入数据 $\boldsymbol{H}^{L-1}$ 转化为量子态。换句话说，就是将第 $L-1$ 个 Transformer 的输出数据编码为量子态。采用角度量子编码方法，将 $\boldsymbol{H}^{L-1}$ 映射到量子比特的旋转角度 $\boldsymbol{\theta}_{ij}$。

$$\boldsymbol{\theta}_{ij}=\pi\times\boldsymbol{h}_{ij}^{L-1} \tag{2}$$

其中，$\boldsymbol{h}_{ij}^{L-1}$ 表示第 $L-1$ 个 Transformer 输出的第 i 个序列的第 j 维特征。每个量子比特通过 Pauli-X 旋转门 $\boldsymbol{R}_x(\theta)$ 实现编码，旋转门的具体计算式如下所示：

$$\boldsymbol{R}_x(\theta)=\mathrm{e}^{-i\theta X/2}=\begin{bmatrix}\cos(\theta/2) & -i\sin(\theta/2)\\ -i\sin(\theta/2) & \cos(\theta/2)\end{bmatrix} \tag{3}$$

其中，$\boldsymbol{X}=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$ 是 Pauli-X 矩阵，表示旋转角度；$\boldsymbol{R}_x(\theta)$ 由 Pauli-X 矩阵作为生成元生成。因此，将输入数据转化为量子态 $|y_i\rangle$ 的计算式如下所示：

$$|y_i\rangle=U_{\mathrm{enc}}(h_i)|0\rangle^{\otimes n} \tag{4}$$

$$U_{\mathrm{enc}}(h_i)=\prod_{i=1}^{n}R_x(h_i,\theta_{ij}) \tag{5}$$

其中，$U_{\mathrm{enc}}(h_i)$ 表示对输入数据的编码操作，$|0\rangle^{\otimes n}$ 表示 n 个量子比特的初始状态为 $|0\rangle$。

基于参数化量子线路操作，初始化 3 个量子电路，其中参

数 q,k,v 分别表示多头注意力中的查询、键和值。对于每个输入状态，分别用 $\langle t_q\rangle_i$，$\langle t_k\rangle_i$ 和 $\langle t_v\rangle_i$ 表示 Pauli-Z1 门的测量输出，具体计算式如下：

$$\langle t_q\rangle_i := \langle y_i | U_q^{\dagger}(\theta_q) Z U_q(\theta_q) | y_i\rangle \tag{6}$$

$$\langle t_k\rangle_i := \langle y_i | U_k^{\dagger}(\theta_k) Z U_k(\theta_k) | y_i\rangle \tag{7}$$

$$\langle t_v\rangle_i := \langle y_i | U_v^{\dagger}(\theta_v) Z U_v(\theta_v) | y_i\rangle \tag{8}$$

将 Pauli-Z1 门的测量输出转化为多头注意力机制中的查询、键和值，分别表示为 $\boldsymbol{T}_q=[\langle t_q\rangle_1,\langle t_q\rangle_2,\cdots,\langle t_q\rangle_n]^{\mathrm{T}}$，$\boldsymbol{T}_k=[\langle t_k\rangle_1,\langle t_k\rangle_2,\cdots,\langle t_k\rangle_n]^{\mathrm{T}}$ 和 $\boldsymbol{T}_v=[\langle t_v\rangle_1,\langle t_v\rangle_2,\cdots,\langle t_v\rangle_n^{\mathrm{T}}$。多头注意力计算式为：

$$A(\boldsymbol{T}_q,\boldsymbol{T}_k,\boldsymbol{T}_v)=\mathrm{softmax}\left(\frac{\boldsymbol{T}_q\boldsymbol{T}_k^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{T}_v \tag{9}$$

$$M(\boldsymbol{T}_q,T_k,T_v)=Concat(\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_n)\boldsymbol{W} \tag{10}$$

其中，A 表示每个头的注意力计算，$\boldsymbol{W}$ 表示权重矩阵。此外，在 Transformer 中通过残差网络和归一化操作，缓解梯度爆炸问题，具体计算式如下所示：

$$\boldsymbol{Z}^{l-1}=LayerNorm(\boldsymbol{H}^0+\boldsymbol{M}^{l-1}) \tag{11}$$

$$FNN(\boldsymbol{Z}^l)=f(\boldsymbol{Z}^{l-1})\boldsymbol{W}^l+\boldsymbol{b}^l \tag{12}$$

其中，第 $l-1$ 层神经元的多头注意力机制输出结果为 $\boldsymbol{Z}^{l-1}$，$\boldsymbol{b}^l$ 表示第 $l-1$ 层到第 l 层的偏差，$\boldsymbol{W}^l$ 表示第 $l-1$ 层到第 l 层的权重矩阵，$FNN(\boldsymbol{Z}^l)$ 表示第 l 层的前馈神经网络输出。

3.3 金字塔视觉特征提取

探索利用图像中视觉对象特征来增强文本实体关系抽取准确率，主要考虑与句子实体直接关联的视觉特征，以及能够表达更加抽象概念的全局图像特征，为文本实体关系抽象任务提供更多的语义知识。具体地，首先采用 Zhang 等[23]提出的视觉对象检测法，提取图像中前 m 个显著性的局部视觉对象；然后，将全局图像 $\boldsymbol{I}$ 和对象图像 $\boldsymbol{O}=\{o_1,o_2,\cdots,o_n\}$ 的尺寸统一缩放到 224×224 像素大小。

在计算机视觉领域，将预训练模型中不同模块的输出特征进行融合已被证明能有效提升模型性能[24]。受此启发，本文重点探索如何在多模态任务中合理利用金字塔特征，提出了一种将图像特征融合到 Transformer 的计算框架中，以增强文本表示能力。具体地，针对给定的图像，采用 CNN 对其进行编码，并通过卷积层块的输出生成一个金字塔特征映射列表，然后通过映射函数将不同尺寸的特征转化为相同尺寸，以匹配 Transformer 计算层的嵌入大小。具体映射过程如下所示：

$$\boldsymbol{V}_i=Conv_{1\times1}(Pool(\boldsymbol{F}_i)),i=1,2,\cdots,c-1 \tag{13}$$

$$\boldsymbol{V}_c=Conv_{1\times1}(\boldsymbol{F}_c) \tag{14}$$

其中，i 和 c 分别表示 CNN 模型中的第 i 个模块和划分的总模块数；$pool()$ 表示池化操作，其目的是将不同尺寸的特征图像映射到固定大小的向量空间中。此外，为了使图像特征与 Transformer 计算层中的文本嵌入具有相同的大小以便进行有效结合，采用了特定的 1×1 卷积层来映射出金字塔形状的视觉特征。

3.4 分层视觉特征融合

如何将这些多尺度的视觉特征与文本特征在 Transformer 框架中进行融合仍是一个挑战。为解决这一难题，首先将分层的多尺度视觉特征进行融合处理，随后将融合后的特征直接与 BERT 模型的第一个 Transformer 层相连。如此，所有的后续 Transformer 层都能够学习到这些经过融合处理的视觉对象特征，从而更有效地结合视觉和文本信息。

为了生成图像特征并确保它们能够与文本特征有效融合，对 Transformer 框架进行了改进：

1）特征标准化：对从不同层级提取的特征 $\boldsymbol{V}_i$ 应用平均池化操作 P，确保每个模块输出的特征具有统一的尺寸 $P(\boldsymbol{V}_i)$。

2）特征融合：为了充分利用不同层级的特征信息，将从 CNN 各个模块中得到的特征进行相加操作，以生成一个综合了多层次信息的平均特征向量。

3）维度调整：为了使得图像视觉特征与 Transformer 的嵌入大小相匹配，通过一个多层感知机（MLP）来调整平均特征向量维度至 c。最终生成的平均特征向量 $\boldsymbol{V}_{\mathrm{output}}$ 计算式如下所示：

$$\boldsymbol{V}_{\mathrm{output}}=\boldsymbol{W}\left(\frac{1}{c}\sum_i^c P(\boldsymbol{V}_i)\right) \tag{15}$$

基于获取的平均特征向量 $\boldsymbol{V}_{\mathrm{output}}$，对最后的聚合层视觉特征进行计算，以匹配第 l 层 Transformer，具体计算式如下所示：

$$\boldsymbol{V}_{\mathrm{output}}^l=\boldsymbol{V}_{\mathrm{output}}\boldsymbol{V}^l \tag{16}$$

图像特征是由视觉全局特征和对象特征所构成。因此，输入到第 l 层 Transformer 的图像特征的具体计算过程如下所示：

$$\hat{\boldsymbol{V}}_{\mathrm{output}}^l=[\hat{\boldsymbol{V}}_{\mathrm{output}}^{(l,I)},\hat{\boldsymbol{V}}_{\mathrm{output}}^{(l,o_1)},\cdots,\hat{\boldsymbol{V}}_{\mathrm{output}}^{(l,o_m)}] \tag{17}$$

3.5 视觉前缀引导融合文本特征

由 3.2 节可知，构建的量子 Transformer 模型可以有效增强模型对文本的理解深度和语义表征能力。此外，为了利用图像信息有效提高文本实体识别和关系抽取的准确率，本文提出一种视觉前缀引导融合文本特征的方法，实现文本与图像间的深度语义交互。视觉前缀引导融合文本特征示意图如图 4 所示。

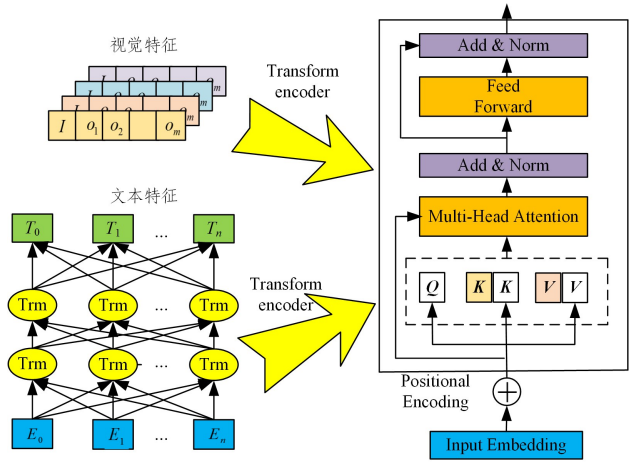


图 4 视觉前缀引导融合文本特征框架

Fig. 4 Framework of visual prefix guided fusion text feature

基于 BERT 的词嵌入：假设给定的输入句子表示为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$，$x_i\in R^{n\times d}$。如图 4 所示，采用 Devlin 等[2]提出的 BERT 模型捕获包含深层语义信息的词汇嵌入序列。具体的计算式如下所示：

$$\begin{aligned}\boldsymbol{H}^l&=\{h_1^l,h_2^l,\cdots,h_n^l\}\\&=BERT(\{x_1,x_2,\cdots,x_n\}),x_i\in\mathbb{R}^{n\times d}\end{aligned} \tag{18}$$

其中，$\boldsymbol{H}^l$ 表示输入序列经过第 l 层 Transformer 的嵌入表示，d 表示隐藏层输出的特征向量维度，n 表示输入序列的长度。

最后，将通过 BERT 生成的文本表示映射到多头注意机制的查询（Query）向量、键（Key）向量和值（Value）向量中。

$$\boldsymbol{Q}^l=\boldsymbol{H}^{l-1}\boldsymbol{W}_l^Q,\boldsymbol{K}^l=\boldsymbol{H}^{l-1}\boldsymbol{W}_l^K,\boldsymbol{V}^l=\boldsymbol{H}^{l-1}\boldsymbol{W}_l^V \tag{19}$$

跨模态特征融合:为了探索文本与图像之间的直接交互,本文采用了多头跨模态注意力机制,将匹配后的对象级图像特征作为键和值,进行注意力计算,得到融合了文本信息的图片表示。具体地,针对聚合的分层视觉特征$\overset{\wedge}{\boldsymbol{V}}{}^{l}_{\text{output}}$。首先,采用一组线性变换操作$\boldsymbol{W}_l^{\varphi}\in\mathbb{R}^{d\times 2\times d}$,将视觉特征映射到与文本表示相同的嵌入空间中;其次,将视觉特征映射为 Transformer 中对应的键向量$\boldsymbol{\varphi}_k^l\in\mathbb{R}^{hw(m+1)\times d}$和值向量$\boldsymbol{\varphi}_v^l\in\mathbb{R}^{hw(m+1)\times d}$。具体的计算式如式(20)所示:

$$\boldsymbol{\varphi}_k^l,\boldsymbol{\varphi}_v^l=\overset{\wedge}{\boldsymbol{V}}{}^{l}_{\text{output}}\boldsymbol{W}_l^{\varphi} \tag{20}$$

其中,$hw(m+1)$表示视觉特征序列的总长度,m 表示采用对象检测算法检测到的视觉对象的数量。融入视觉特征的 Transformer 计算层的具体结构如图 4 所示。

$$Attention^l=\text{softmax}\left(\frac{\boldsymbol{Q}^l[\boldsymbol{\varphi}_k^l;\boldsymbol{K}^l]}{\sqrt{d}}\right)[\boldsymbol{\varphi}_v^l;\boldsymbol{V}^l]^{\mathrm{T}} \tag{21}$$

将不同尺度的视觉特征直接输入到 BERT 模型的第一个 Transformer 层中(详见 3.2 节),可使所有后续的 Transformer 层都学习到这些视觉特征所携带的视觉对象信息。模型结合文本上下文信息与图像语义信息,此方法能够有效地编码两者的信息,从而有助于减少不相关的图像数据对整体模型性能的负面影响。

3.6 实体关系标注

基于以上的论述,获得 BERT 的最终输出向量$\boldsymbol{H}^L=\boldsymbol{U}(\boldsymbol{X},\overset{\wedge}{\boldsymbol{V}}{}^{L}_{\text{output}})$。其中$U(\cdot)$函数表示视觉特征与文本特征的连接操作。最后,针对 MNER 和 MRE 两个任务,分别设计不同的标注模型。

针对多模态实体识别任务。利用条件随机场(CRF)对输入序列进行标注。该方法不仅能够充分利用相邻标签间的相关性,还能对整条标签序列进行评分。具体而言,针对输入序列中的每个词,给出对应的标签编码。

$$P(y|\boldsymbol{H}^L)=\frac{\prod_{i=1}^{n}S_i(y_{i-1},y_i\boldsymbol{H}^L)}{\sum_{y'\in Y}\prod_{i=1}^{n}S_i(y'_{i-1},y_i{}'\boldsymbol{H}^L)} \tag{22}$$

$$L_{\text{NER}}=-\sum_{i=1}^{M}\log(P(y^i|U(\boldsymbol{X}^i,\overset{\wedge}{\boldsymbol{V}}_{output}))) \tag{23}$$

其中,Y 表示输入句子对应的 BIO(B-begin,I-inside,O-outside)预测标签序列集合;$S(\cdot)$表示对应标签的得分函数。L_{NER}是训练过程中采用最大似然函数对输入序列进行标注预测的结果。

多模态关系抽取任务的其主要目标是预测主体实体和对象实体之间的语义关系$r\in Y$。形式上,假设输入序列中存在两个实体$X=\{x_1,x_2,\cdots,x_n\}$,分别为$E_1=\{x_i,\cdots,x_{i+|E_1|-1}\}$和$E_1=\{x_i,\cdots,x_{i+|E_2|-1}\}$,关系抽取任务可以视作一个分类任务,根据预定义的关系类型集 Y 来确定实体E_1和实体E_2之间的语义关系类型。采用 Softmax()函数计算实体间关系的概率分布,具体如下所示:

$$P(r|X,E_1,E_2)=\text{softmax}(\boldsymbol{W}([E_1\oplus E_2])+\boldsymbol{b}) \tag{24}$$

其中,$\boldsymbol{W}$表示可学习权重,$\boldsymbol{b}$表示偏置向量。最后,通过交叉熵损失函数来计算关系抽取模型的损失,具体计算式如下所示:

$$L_{\text{MRE}}(P(y|X))=-\sum_{i}^{M}\log(P(r|X,E_1,E_2)) \tag{25}$$

4 实验分析

4.1 实验设置

本文将两个公开的多模态实体和一个关系抽取数据集作为数据基础。Twitter-2015 和 Twitter-2017 数据集包含了文本及其对应的图片,这些数据集不仅标注了目标实体,还标注了图文中表达的情感倾向(见表 1)。由 Zheng 等[25]构建的数据集则包含超过 15000 个实例和 23 种预定义的关系类型(见表 2),用于多模态关系抽取任务。

表 1 多模态实体识别数据集

Table 1 Multimodal entity recognition datasets

实体	Twitter-2015			Twitter-2017		
Per	2217	552	1816	2943	626	621
Loc	2091	522	1697	731	173	178
Org	928	247	839	1674	375	395
Misc	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351

表 2 多模态关系抽取数据集

Table 2 Multimodal relation extraction datasets

数据集	单词	句子	实例	实体	关系	图像
SemEval-2010	205000	10717	8853	21434	9	0
MNER	258000	**9201**	15485	30970	23	**9201**

本文采用精确率(Precision)、召回率(Recall)和 F1 分数来评估模型性能。F1 值是精确度和召回率的调和平均,平衡了模型的准确性和完整性。

$$P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN},F1=\frac{2PR}{P+R} \tag{26}$$

其中,TP 代表模型中正确地识别为实体或关系的数量;FP 是指模型中错误地标记为实体或关系的数量;而 FN 则表示实际上属于实体或关系但模型未能正确识别的数量。

在模型训练的过程中,本文为 MNER 模型和 MRE 模型挑选了最优的超参数配置,具体参数值如表 3 所列。

表 3 模型参数设置

Table 3 Model parameter settings

参数名	Twitter-2015	Twitter-2017	MNRE
最大句子长度	256	256	256
Batch 大小	16	16	32
BERT 隐藏层维度	768	768	768
BERT 隐藏层个数	12	12	12
Dropout	0.5	0.5	0.4
学习率	3×10^{-5}	3×10^{-5}	1×10^{-5}
迭代次数	10	10	15
视觉对象 m	4	4	4

在实验过程中,采用了 PyTorch 框架作为基本的开发平台,并使用 Python3 作为编程语言。此外,所提出模型以及基准模型的训练和测试主要是在 P100 GPU 上进行的。

4.2 实验结果与分析

为了验证所提出模型在处理多模态信息方面的优势,同时确保模型在不同数据集上的一致表现,在 Twitter-2015,Twitter-2017 以及 MNRE 这 3 个数据集上进行了对比实验。

基于文本的单模态模型:该类基线模型主要包含 CNN-BLSTM-CRF,HBiLSTM-CRF 和 BERT-CRF 等经典序列标注方法,这些模型在新闻领域的命名实体识别(NER)任务中展现出了优异的预测能力。此外,脉冲耦合神经网络(PCNN)[26]作为一种利用外部知识库信息来进行关系抽取的远程监督方法,也在该领域得到了应用。而多任务双向编码器(MTB)[27]则代表了一种先进的技术手段,被广泛应用于多种基于文本的关系抽取(RE)任务中。

基于多模态的模型：AdapCoAtt[14]是最早利用共享注意力机制完成多模态命名实体识别任务的模型之一。另一方面，RpBERT[5]通过在 Twitter 数据上训练分类器，能够精确地计算图像与文本之间的相似性。此外，OCSGA[28]，UMT[16]，UMGF[25]和 MEGA[4]均为多模态实体识别和关系抽取方法，它们通过使用 Transformer 或图神经网络（GNN）将视觉对象特征与文本表示对齐，从而增强模型的多模态理解能力。VisualBERT[18]则是一种视觉预训练语言模型，适用于多模态命名实体识别（MNER）和多模态关系抽取（MRE）任务，它通过整合视觉信息与语言信息来改进模型性能。更进一步地，HVPNet[29]，MEGA[30]和 TMR[31]代表了从多个视角学习视觉前缀的层次结构的方法，这些模型通过多层次的视觉信息处理，在多模态命名实体识别和关系抽取任务中达到了较为先进的水平。

将提出的基于量子 Transformer 的多模态实体关系联合抽取模型与基线模型进行了性能对比。具体结果如表 4—表 6 所列。

表 4 与基线模型在 Twitter-2015 数据集上的性能比较

Table 4 Performance comparison with baseline model on Twitter-2015 dataset

（%）

类型	方法	*P*	*R*	*F*1
Text	CNN-BiLSTM-CRF	66.24	68.09	67.15
	HBiSTM-CRF	70.32	68.05	69.17
	BERT-CRF	69.22	74.59	71.81
	PCNN	N/A	N/A	N/A
	MTB	N/A	N/A	N/A
Text + Image	AdapCoAtt	69.87	74.59	72.15
	OCSGA	74.71	71.21	72.92
	RpBERT	71.15	74.30	72.69
	UMT	71.67	75.23	73.41
	UMGF	74.49	75.21	74.85
	VisualBERT	68.84	71.39	70.09
	MEGA	70.35	74.58	72.35
	HVPNeT	73.87	76.82	75.32
	TMR	**75.26**	76.49	75.87
	Ours	75.12	**77.08**	**76.09**

表 5 与基线模型在 Twitter-2017 数据集上的性能比较

Table 5 Performance comparison with baseline model on Twitter-2017 Datasets

（%）

类型	方法	*P*	*R*	*F*1
Text	CNN-BiLSTM-CRF	80.00	78.76	79.37
	HBiSTM-CRF	82.69	78.16	80.37
	BERT-CRF	83.32	83.57	83.44
	PCNN	N/A	N/A	N/A
	MTB	N/A	N/A	N/A
Text + Image	AdapCoAtt	85.13	83.20	84.10
	OCSGA	N/A	N/A	N/A
	RpBERT	N/A	N/A	N/A
	UMT	85.28	85.34	85.31
	UMGF	86.54	84.50	85.51
	VisualBERT	84.06	85.39	84.72
	MEGA	84.03	84.75	84.39
	HVPNeT	85.84	87.93	86.87
	TMR	**88.12**	88.38	88.25
	Ours	87.76	**89.15**	**88.45**

表 6 与基线模型在 MNRE 数据集上的性能比较

Table 6 Performance comparison with baseline model on MNRE Datasets

（%）

类型	方法	*P*	*R*	*F*1
Text	CNN-BiLSTM-CRF	N/A	N/A	N/A
	HBiSTM-CRF	N/A	N/A	N/A
	BERT-CRF	N/A	N/A	N/A
	PCNN	62.85	49.69	55.49
	MTB	64.46	57.81	60.86
Text + Image	AdapCoAtt	N/A	N/A	N/A
	OCSGA	N/A	N/A	N/A
	RpBERT	N/A	N/A	N/A
	UMT	62.93	63.88	63.46
	UMGF	64.38	66.23	65.29
	VisualBERT	57.15	59.48	58.30
	MEGA	64.51	68.44	66.41
	HVPNeT	83.64	80.78	81.85
	TMR	90.48	**87.66**	89.05
	Ours	**91.05**	87.22	**89.09**

实验结果表明，在 Twitter-2015，Twitter-2017 和 MNRE 这 3 个测试集上，本文提出的多模态命名实体识别（MNER）和多模态关系抽取（MRE）模型均表现出优于基线方法的性能。这些结果显示了所提模型在处理多模态数据时的强大能力，证明了其在不同数据集上的一致优越性。

首先，将提出的多模态实体关系抽取模型与基线方法进行了对比。结果显示，融合了视觉特征的方法明显优于仅依赖文本特征的实体识别方法。例如，UMGF 相比于 NERT-CRF，在实体识别任务上 F1 值提升了约 2 个百分点；而 MEGA 相较于 MTB，在关系抽取任务上的 F1 值则提高了约 5 个百分点。这表明，引入视觉特征有助于增强文本的语义表示，进而显著提升实体识别与关系抽取模型的性能。

其次，提出的方法在性能上超越了最新的 HVPNeT 和 TMR 方法。具体而言，在 Twitter-2017 和 MNRE 数据集上，本文方法在 F1 值上都高于基线模型。更重要的是，以往的多模态实体识别和关系抽取方法未充分考虑到图片中无关对象特征所带来的误差影响。相比之下，本文提出的方法通过将视觉对象特征作为文本的提示信息，从而增强了文本的语义表示。

最后，将本文提出的方法与 VisualBERT 进行了比较。VisualBERT 是一种用于处理多模态数据的 BERT 改进模型。尽管该模型在多模态实体识别和关系抽取任务中有效提升了性能，但其表现仍不及 UMGF 和 MEGA 模型，这表明 VisualBERT 在某些方面仍有待改进。导致这一结果的主要原因可能是数据集与预训练过程中所学知识之间的相关性较弱，从而影响了模型在多模态任务中的表现。

4.3 基于消融实验的模型性能分析

为了评估提出的多模态实体关系联合抽取模型中各模块的贡献，在 3 个数据集上进行了相应的消融实验，结果如表 7—表 9 所列。

表 7—表 9 中的结果表明本文提出的模型具有更好的性能。将量子计算融合到 Transformer 框架中，模型的性能得到了显著的提高。此外，通过使用金字塔视觉特征提取模型，能够捕获从高层到底层的金字塔状层次特征，从而有效提高了模型的性能。

表 7　在 Twitter-2015 数据集上模型中各个模块的影响

Table 7　Impact of each module in the model on Twitter-2015 datasets

（%）

模型	*P*	*R*	*F*1
w/o 量子模块和金字塔模块	69.25	72.02	70.61
w/o 量子模块	73.68	75.92	74.78
w/o 金字塔模块	74.35	76.14	75.23
Ours	75.12	77.08	**76.09**

表 8　在 Twitter-2017 数据集上模型中各个模块的影响

Table 8　Impact of each module in the model on Twitter-2017 Datasets

（%）

模型	*P*	*R*	*F*1
w/o 量子模块和金字塔模块	83.64	84.24	83.94
w/o 量子模块	85.14	87.59	86.35
w/o 金字塔模块	85.69	87.85	86.76
Ours	87.76	89.15	**88.45**

表 9　在 MNRE 数据集上模型中各个模块的影响

Table 9　Impact of each module in the model on MNRE dataset

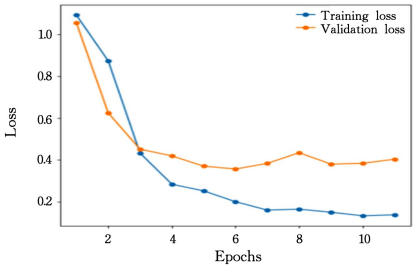
（%）

模型	*P*	*R*	*F*1
w/o 量子模块和金字塔模块	80.53	78.81	79.66
w/o 量子模块	88.20	84.72	86.42
w/o 金字塔模块	89.38	86.12	87.72
Ours	91.05	87.22	**89.09**

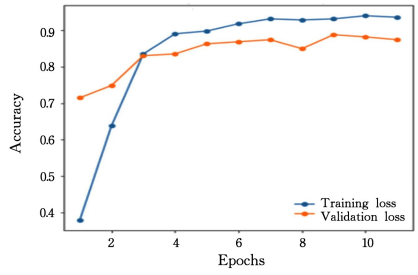
4.4　基于损失与准确率曲线的模型性能分析

为了进一步证明提出的多模态实体关系抽取模型是有效且稳定的，本文以 Twitter-2017 数据集为数据基础，探索和分析模型在训练集和测试集上的损失曲线和准确率曲线。

如图 5 所示，损失曲线随着训练轮数的增加而平稳下降，迭代到 10 次后趋向平稳，表明模型在训练集和测试集上的拟合情况良好。同时，准确率曲线随着迭代次数增加而逐渐上升，最终在模型迭代 10 次时趋向平稳，表明模型在训练集和测试集上的性能越来越好。此外，损失曲线稳定下降至低点，同时准确率曲线稳步上升至高点，有效说明本文模型具有较好的泛化能力。



（a）Training and validation loss



（b）Training and validation accuracy

图 5　模型在 Twitter-2017 数据集上的损失和准确率曲线

Fig. 5　Mdel loss and accuracy curve on Twitter-2017 dataset

结束语　本文提出了一种基于量子 Transformer 的多模态实体关系联合抽取方法。首先，设计一种针对文本数据处理的参数化量子电路，实现文本数据深层特征的提取；其次，通过设计的金字塔视觉特征提取模型获取包含从高到底的金字塔状的层次特征，充分考虑了图像的多尺度信息；最后，通过设计的分层视觉前缀网络将分层多尺度图像特征与文本特征对齐并融合，获取鲁棒性高的文本表示。实验结果表明，提出的多模态实体关系联合抽取模型在性能上是有效且稳定的。

尽管目前提出的 MNER 和 MRE 方法已经取得了显著的进展，但当图像内容与文本信息不匹配时，这两种模型的表现依然较差。鉴于此，后续的研究将致力于研究动态过滤图像中的潜在噪声，从而增强模型对非一致性模态信息的鲁棒性。此外，由于现有关于 MNER 和 MRE 任务的数据集较少，这在一定程度上阻碍了相关研究的发展。因此，下一步将利用不同社交平台中大量未标记的社交帖子，结合少量的标注数据，训练更强大的 MNER 模型和 MRE 模型。

参 考 文 献

[1] LI J，SUN A，HAN J，et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering，2020，34(1)：50-70.

[2] LI D，YAN L，YANG J，et al. Dependency syntax guidedbert-bilstm-gam-crf for chinese ner[J]. Expert Systems with Applications，2022，196：116682.

[3] MOON S，NEVES L，CARVALHO V. Multimodal Named Entity Recognition for Short Social Media Posts[C]// Proceedings of NAACL-HLT. 2018：852-860.

[4] ZHENG C，FENG J，FU Z，et al. Multimodal relation extraction with efficient graph alignment[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021：5298-5306.

[5] SUN L，WANG J，ZHANG K，et al. RpBERT：a text-image relation propagation-based BERT model for multimodal NER[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021：13860-13868.

[6] XU Z，WANG C，QIU M，et al. Making pre-trained language models end-to-end few-shot learners with contrastive prompt tuning[C]// Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023：438-446.

[7] SUN S，GAO H. Meta-AdaM：An meta-learned adaptive optimizer with momentum for few-shot learning[J]. Advances in Neural Information Processing Systems，2023，36：65441-65455.

[8] WANG Y，SUN Y，FU Y，et al. Spectrum-BERT：pre-training of deep bidirectional transformers for spectral classification of Chinese liquors[J]. IEEE Transactions on Instrumentation and Measurement，2024，73：1-13.

[9] HAN B，HE L，KE J，et al. Weighted parallel decoupled feature pyramid network for object detection[J]. Neurocomputing，2024，593：127809.

[10] TIWARI P，ZHANG L，QU Z，et al. Quantum fuzzy neural network for multimodal sentiment and sarcasm detection[J]. Information Fusion，2024，103：102085.

[11] PHUKAN A，HAQ KHAN A，EKBAL A. QuMIN：quantum multi-modal data fusion for humor detection[J]. Multimedia Tools and Applications，2025，84(18)：18855-18872.

[12] XU B，HUANG S，SHA C，et al. MAF：a general matching and

alignment framework for multimodal named entity recognition [C]// Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022:1215-1223.
[13] CHEN X, ZHANG N, XIE X, et al. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction[C]// Proceedings of the ACM Web Conference 2022. 2022:2778-2788.
[14] ZHANG Q, FU J, LIU X, et al. Adaptive co-attention network for named entity recognition in tweets[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.
[15] NIE Y, TIAN Y, WAN X, et al. Named Entity Recognition for Social Media Texts with Semantic Augmentation[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2020:1383-1391.
[16] YU J, JIANG J, YANG L, et al. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3342-3352.
[17] LI G, DUAN N, FANG Y, et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:11336-11344.
[18] LI L H, YATSKAR M, YIN D, et al. Visualbert: A simple and performant baseline for vision and language[J]. arXiv: 1908. 03557, 2019.
[19] SU W, ZHU X, CAO Y, et al. VL-BERT: Pre-training of Generic Visual-Linguistic Representations[C]// International Conference on Learning Representations. 2019.
[20] CHEN Y C, LI L, YU L, et al. Uniter: Universal image-text representation learning[C] // European Conference on Computer Vision. Cham: Springer International Publishing, 2020:104-120.
[21] TAN H, BANSAL M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019:5100-5111.
[22] LU J, BATRA D. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. Advances in Neural Information Processing Systems, 2019, 32.
[23] ZHANG D, WEI S, LI S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021:14347-14355.
[24] ZHANG T M, ZHANG S, LIU X, et al. Multimodal Data fusion for Few-shot Named Entity Recognition Method[J]. Journal of Software, 2024, 35(3):1107-1124.
[25] ZHENG C, FENG J, FU Z, et al. Multimodal relation extraction with efficient graph alignment[C] // Proceedings of the 29th ACM International Conference on Multimedia. 2021:5298-5306.
[26] WU J K, LI W J. Remote Supervised Relationship Extraction Method Based on PCNN Similar Sentence Bag Attention [J]. Journal of Chinese Information Science, 2024, 38(5):65-75.
[27] SOARES L B, FITZGERALD N, LING J, et al. Matching the Blanks: Distributional Similarity for Relation Learning[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:2895-2905.
[28] WU Z, ZHENG C, CAI Y, et al. Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020:1038-1046.
[29] CHEN X, ZHANG N, LI L, et al. Good Visual Guidance Make A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction[C]// Findings of the Association for Computational Linguistics: NAACL 2022. 2022:1607-1618.
[30] CHEN X, ZHANG N, LI L, et al. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion [C]// Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022:904-915.
[31] ZHENG C, FENG J, CAI Y, et al. Rethinking Multimodal Entity and Relation Extraction from a Translation Point of View[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023:6810-6824.

LI Daiyi, born in 1988, Ph.D, is a member of CCF(No. 58208G). His main research interests include natural language processing, knowledge graphs and big data, etc.

WU Huaiguang, born in 1976, Ph. D, professor, is a member of CCF (No. 13128D). His main research interests include big data, ubiquitous computing, and formal methods, etc.