计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于多任务学习的多模态命名实体识别方法
作者：李晓腾，张盼盼，勾智楠，高凯
DOI：10.19678/j.issn.1000-3428.0064087
网络首发日期：2022-06-21
引用格式：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[J/OL]．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0064087

网络首发时间：2022-06-21 11:56:40
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.TP.20220620.1211.003.html

计 算 机 工 程
Computer Engineering

基于多任务学习的多模态命名实体识别方法

李晓腾[1]，张盼盼[1]，勾智楠[2]，高凯[1]
（1.河北科技大学，信息科学与工程学院，河北石家庄 050018；2.河北经贸大学，信息技术学院，河北石家庄 050061）

摘 要：针对传统多模态命名实体识别方法无法有效融合图文模态信息且无法有效区分易混淆实体等问题，提出一种基于多任务学习的多模态命名实体识别方法。该方法通过对比融合辅助任务来促进图文模态信息的融合，并通过实体聚类辅助任务来提升模型对易混淆实体的判断能力。首先，利用 Bert 预训练语言模型和 ResNet 模型分别对原始文本和图片进行特征映射获得相应的特征向量，并利用跨模态 Transformer 结构融合图文模态信息。其次，在多模态命名实体识别任务基础上，增加对比融合辅助任务促进图文模态信息融合；增加实体聚类辅助任务来学习实体类别之间的差异，提升模型对易混淆实体的区分能力。最后，利用 CRF 层学习上下文转移概率，并输出最优预测结果。实验结果表明，在国际公开数据集 Twitter-2017 上，相较于基线方法所提方法取得更高的准确率、召回率和 F1 值，其中 F1 值可达 85.59%。说明了所提方法对于多模态命名实体识别任务是有效的，对比融合辅助任务和实体聚类辅助任务促进了模型对实体的识别效果。

关键词：命名实体识别；多任务学习；多模态；对比学习；聚类

开放科学（资源服务）标志码（OSID）：

Multimodal Named Entity Recognition based on Multi-Task Learning

LI Xiaoteng[1]，ZHANG Panpan[1]，GOU Zhinan[2]，GAO Kai[1]

(1. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China;
2. School of Information Technology, Hebei University of Economics and Business, Shijiazhuang, Hebei, 050061, China)

【Abstract】Aiming at the problem that traditional multi-mode named entity recognition methods cannot effectively integrate the modal information of text and image and cannot effectively distinguish confusable entities, a multi-mode named entity recognition method based on multi-task learning is proposed. In this method, the modal information fusion is promoted by contrast and fusion auxiliary task, and the judgment ability of confusable entities is improved by entity clustering auxiliary task. Firstly, Bert pre-trained language model and ResNet model are used to obtain the feature vectors, and cross-modal Transformer is used to fuse the modal of text and image. Secondly, on the basis of the multi-modal named entity recognition task, the auxiliary task of contrast fusion is added to promote the image and text modal information fusion. The auxiliary task of entity clustering is added to learn the differences between entity categories and improve the ability of the model to distinguish easily confused entities. Finally, CRF layer is used to learn the context transition probability and output the optimal prediction results. Experimental results show that in the international open dataset Twitter-2017, the method we proposed achieves higher accuracy, recall rate and F1 value than the baseline methods, in which the F1 value can reach 85.59%. It is shown that the proposed method is effective for multi-modal named entity recognition task, and the comparison fusion auxiliary task and entity clustering auxiliary task can improve the recognition effect of the model.

【Key words】named entity recognition; multi-task learning; multi-modal; contrastive learning; clustering
DOI:10.19678/j.issn.1000-3428.0064087

0 概述

命名实体识别（named entity recognition, NER）是指抽取文本序列中的“人名”、“地名”、“机构名”等实体，是一项重要的自然语言处理任务。命名实体识别任务广泛应用于其他自然语言处理任务，如信息抽取、信息检索、问答系统以及构建知识图谱等[1]。随着社交媒体网络的快速发展，浩如烟海的多模态社交网络数据亟待处理。多模态命名实体识别任务需要在一段文本序列 S 及对应的图片 V 中，判断出文本序列中的实体，并对这些实体分类。MOON 等[2]使用双向长短时记忆网络（bi-directional long-short term memory，Bi-LSTM）和条件随机场（conditional random field，CRF）为基础模型结构，CNN 模块抽

基金项目：河北省高等学校科学技术研究项目资助(QN2020198)；河北省自然科学基金（预研）（F2022208006）
作者简介：李晓腾（1994—），男，河北石家庄人，硕士研究生，主要从事自然语言处理方面的研究；张盼盼，硕士研究生；勾智楠，博士，讲师；高凯（通信作者），硕士生导师，教授。**E-mail：**gaokai@hebust.edu.cn

取图像特征，并利用注意力机制为各类特征计算权重。ZHANG 等[3]以 Bi-LSTM+CRF 为基本框架，使用 VGGNET-16 抽取图片特征，并通过互注意力层计算融合权重，融合特征通过 CRF 获取预测结果。LU 等[4]以 Bi-LSTM+CRF 为基础模型框架，使用 ResNet[5]抽取图片特征，利用文本特征作为查询向量计算得到相关度高的图片特征，并利用门控机制融合图片特征和文本特征。YU 等[6]利用 ResNet 抽取图片特征，通过预训练模型 BERT[7]获取文本特征表示，利用跨模态 Transformer 结构来融合图文 2 种模态信息。多任务学习广泛应用于图像和自然语言处理任务中[8~11]。多任务学习指多个相关任务联合训练，通过共享任务间信息，帮助主任务学习[12]。多任务学习中参数共享方式有硬共享、软共享等[13~14]。多任务学习在命名实体识别任务中同样有广泛应用，REI 等[15]提出利用无监督辅助任务来帮助网络模型去学习深层的文本语义、语法信息。LIN 等[16]提出一种跨语言多任务学习方法，来缓解特定 NER 领域语料不足的问题。

虽然多模态命名实体识别任务中已有许多优秀的工作，但是仍然存在以下亟待解决的问题：如何有效融合图文 2 种模态信息；如何高效区分易混淆实体。为了解决上述问题，本文提出基于多任务的多模态命名实体识别算法（multi-task learning multimodal named entity recognition, MLMNER）。首先，提出对比融合辅助任务来进一步促进图文 2 种模态信息的融合。对比融合辅助任务旨在拉近同一对样本中图文特征表示在投影空间的距离，以此来保证融合后的图文特征保持较强的相关性，进而提升多模态信息的融合效果。其次，提出实体聚类辅助任务，来学习实体之间类别的差异。实体聚类辅助任务旨在学习实体类别之间的差异，进一步帮助命名实体识别任务学习更好的特征表示，从而提升实体识别效果。

1 联合实体边界检测的命名实体识别模型

1.1 模型结构

本文方法模型结构如图 1 所示，模型整体从左向右分为 3 个子模块，第 1 部分为表示嵌入模块，通过 Bert 和 ResNet 将原始文本和图片映射为特征向量；第 2 部分为特征融合模块，利用跨模态 Transformer 结构融合图文 2 种模态信息；第 3 部分是多任务学习模块，包含 4 个任务。

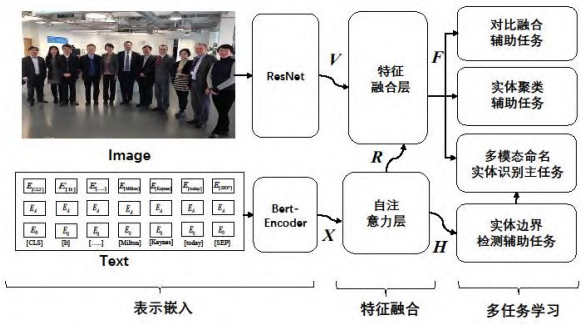

图 1 MLMNEER 模型结构

Fig.1 MLMNER model architecture

1.1.1 表示嵌入

文本嵌入：如图 1 表示嵌入层所示，Bert 模型作为文本编码器。对于输入长度为 n 的文本 S，本文定义 $S'=(s_0,s_1,s_2,...,s_{n+1})$ 为 Bert 编码器的输入，s_0 和 s_{n+1} 分别代表文本开始字符[CLS]和结束字符[SEP]。s_i 由 token 嵌入、segment 嵌入、position 嵌入构成。$\boldsymbol{X}=(\boldsymbol{x}_0,\boldsymbol{x}_1,...,\boldsymbol{x}_{n+1})$ 为 Bert 编码层的输出，$\boldsymbol{x}_i\in\mathbb{R}^d$ 是 s_i 的词特征向量，d 是特征维度。

图片嵌入：如图 1 表示嵌入层所示，ResNet 用来抽取图像特征。I 是文本对应的图片，I 经 ResNet 网络得到的最终特征向量表示为 $\boldsymbol{I}=(\boldsymbol{i}_1,\boldsymbol{i}_2,...,\boldsymbol{i}_{49})$，$\boldsymbol{i}_i\in\mathbb{R}^{2048}$。为了方便后续做模态交互，通过线性层来调整图片向量的维度，$\boldsymbol{V}$ 为图片嵌入输出，$\boldsymbol{V}=(\boldsymbol{v}_1,\boldsymbol{v}_2,...,\boldsymbol{v}_{49})$，$\boldsymbol{v}_i\in\mathbb{R}^d$，$d$ 是特征维度。

1.1.2 特征融合

为丰富文本序列每个词的上下文信息，$\boldsymbol{X}$ 经过一层标准 Transformer 层[17]来捕获上下文信息，得到新的表示 $\boldsymbol{R}=(\boldsymbol{r}_0,\boldsymbol{r}_1,...,\boldsymbol{r}_{n+1})$，其中 $\boldsymbol{r}_i\in\mathbb{R}^d$，$d$ 是特征维度。

为了更好的融合文本和图片 2 种模态信息，引入多头跨模态 Transformer 层。如图 2 所示，多头跨模态 Transformer 层使用另一种模态信息作为查询向量。因此，可以学习到与另一种模态相关的语义信息，促进模态信息融合。多头跨模态 Transformer 可以分为 2 部分：图片指导的文本表示模块和文本指导的图片表示模块。

图片指导的文本表示：利用图片特征向量 $\boldsymbol{V}$ 作为模态交互层的 $\boldsymbol{Q}$ 向量，文本特征向量 $\boldsymbol{R}$ 作为 $\boldsymbol{K}$ 向量和 $\boldsymbol{V}$ 向量。多头跨模态 Transformer 网络具体计算如式(1)~式(2)所示：

$$\mathrm{Att}_i(\boldsymbol{V},\boldsymbol{R})=\mathrm{softmax}\left(\frac{\left[\boldsymbol{W}_{\mathrm{q_i}}\boldsymbol{V}\right]\left[\boldsymbol{W}_{\mathrm{K_i}}\boldsymbol{R}\right]^T}{\sqrt{\mathrm{d}/\mathrm{m}}}\right)\left[\boldsymbol{W}_{\mathrm{V_i}}\boldsymbol{R}\right] \quad (1)$$

$$\mathrm{MultiHeadAtt}(\boldsymbol{V},\boldsymbol{R})=\boldsymbol{W}'[\mathrm{Att}_1(\boldsymbol{V},\boldsymbol{R}),...,\mathrm{Att}_m(\boldsymbol{V},\boldsymbol{R})] \quad (2)$$

式中：$\mathrm{Att}_i(\boldsymbol{V},\boldsymbol{R})$ 是指第 i 个跨模态交互注意力

网络计算公式。MultiHeadAtt($\boldsymbol{V},\boldsymbol{R}$)是指多个跨模态交互注意力网络的拼接。$\boldsymbol{W}_{q_i},\boldsymbol{W}_{k_i},\boldsymbol{W}_{v_i}$ 是权重矩阵，$\boldsymbol{W}'$ 是多头跨模态交互注意力网络的权重矩阵。

随后，经正则化层和前馈网络层得到跨模态交互网络输出 $\boldsymbol{P}\in(\boldsymbol{p}_1,\boldsymbol{p}_2,...,\boldsymbol{p}_{49})$。添加 Transformer 模块使得最终输出特征长度与文本长度保持一致，图片指导的文本表示最终为 $\boldsymbol{T}=(\boldsymbol{t}_0,\boldsymbol{t}_1,...,\boldsymbol{t}_{n+1})$。

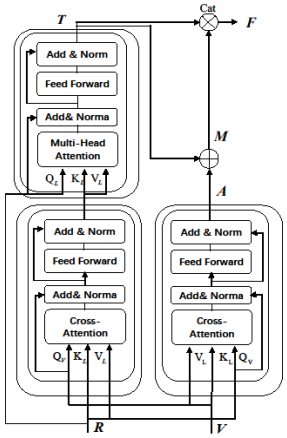

图 2　多头跨模态 Transformer

Fig.2　Multi-Head Cross-Transformer

文本指导的图片表示：利用文本特征向量 $\boldsymbol{R}$ 作为跨模态交互层的 $\boldsymbol{Q}$ 向量，图片特征向量 $\boldsymbol{V}$ 作为 $\boldsymbol{K}$ 向量和 $\boldsymbol{V}$ 向量。最终跨模态交互注意力网络层的输出是 $\boldsymbol{A}=(\boldsymbol{a}_0,\boldsymbol{a}_1,...,\boldsymbol{a}_{n+1})$。文本序列中一些虚词是不必要的，因此引入门控机制控制 2 种模态信息对齐，最终文本指导的图片表示为 $\boldsymbol{M}$。

本文将得到的文本表示 $\boldsymbol{T}$ 和图片表示 $\boldsymbol{M}$ 拼接起来作为融合后的特征表示 $\boldsymbol{F}=(\boldsymbol{f}_0,\boldsymbol{f}_1,...,\boldsymbol{f}_{n+1})$，其中 $\boldsymbol{f}_i\in\mathbb{R}^{2d}$。

1.1.3 多任务学习

针对多模态命名实体识别任务特点，本文添加了 3 个与多模态命名实体识别相关的辅助任务，旨在联合辅助任务帮助模型学习相关的语义知识，提升模型对实体识别能力。首先，为了促进多模态命名实体识别任务中图文模态融合，添加对比融合辅助任务来拉近图文特征向量距离，保证两者保持较强的相关性；其次，为了增强模型对易混淆实体的判断能力，添加实体聚类辅助任务来学习各类实体的独有特征，增强对易混淆实体的判断能力；最后，为增强模型对实体边界信息的利用，添加边界检测任务来学习实体边界信息。下面将具体介绍各个辅助任务以及多模态命名实体识别任务的模型细节。

对比融合辅助任务：对比学习是一种自监督表示学习策略[18]，是将正例样本和负例样本映射到特征空间之后学习两者之间的差异，进而学习到样本的独特特征，代表工作有[19~22]。据此，本文提出对比融合辅助任务来拉近样本中图文特征表示在投影空间的距离，以此来保证融合后的图文特征表示保持较强的相关性。文本向量 $\boldsymbol{T}$ 和图片向量 $\boldsymbol{M}$ 通过对比损失来保持相似性。对比融合损失包含图像到文本的对比损失和文本到图像的对比损失。

同一 batch 中的图文对经过特征融合模块后得到成对特征 $(\bar{\boldsymbol{M}}_i,\bar{\boldsymbol{T}}_i)$，其中 $i\in(1,2,...,N)$，N 是 batch-size 的大小。$\bar{\boldsymbol{M}}_i$ 是图片特征取平均后的表示，$\bar{\boldsymbol{T}}_i$ 是文本特征取平均后的表示。对于第 i 组图文对，由图像到文本的对比融合损失如式(3)所示：

$$l_i^{(\mathrm{M}\to\mathrm{T})}=-\log\frac{\exp\left([\bar{\boldsymbol{M}}_i,\bar{\boldsymbol{T}}_i]/t\right)}{\sum_{k=1}^{N}\exp\left([\bar{\boldsymbol{M}}_i,\bar{\boldsymbol{T}}_k]/t\right)}\tag{3}$$

式中：$[\bar{\boldsymbol{M}}_i,\bar{\boldsymbol{T}}_i]$ 为图片模态与文本模态余弦相似度；t 为温度系数，k 表示同一 batch 中非 i 样本。同理，由文本到图像的对比融合损失如式(4)所示：

$$l_i^{(\mathrm{T}\to\mathrm{M})}=-\log\frac{\exp\left([\bar{\boldsymbol{T}}_i,\bar{\boldsymbol{M}}_i]/t\right)}{\sum_{k=1}^{N}\exp\left([\bar{\boldsymbol{T}}_i,\bar{\boldsymbol{M}}_k]/t\right)}\tag{4}$$

最终，对比融合损失如式(5)所示：

$$\mathrm{loss}_{\mathrm{Fusion}}=-\frac{1}{N}\sum_{j=1}^{N}\lambda\boldsymbol{l}_j^{(\mathrm{M}\to\mathrm{T})}+(1-\lambda)\boldsymbol{l}_j^{(\mathrm{T}\to\mathrm{M})}\tag{5}$$

式中，N 是 batch-size 的大小，λ 是图像到文本对比融合损失的权重系数。

实体聚类辅助任务：多模态命名实体识别数据中存在许多易混淆实体，传统模型往往对这些实体无法有效区分。聚类指的是按照一定的规则将数据进行划分为不同的簇，并且簇内数据尽可能相似，簇之间的数据差异尽可能大。因此，本文借鉴聚类思想引入实体聚类辅助任务来学习每个类的独有特征，进而增强对易混淆实体的区分能力。

首先，本文通过预训练的方式获得各个实体类中心的特征向量表示 $\boldsymbol{C}=(\boldsymbol{c}_1,\boldsymbol{c}_2,\boldsymbol{c}_3,\boldsymbol{c}_4)$，$\boldsymbol{c}_k\in\mathbb{R}^{2d}$，$\boldsymbol{c}_k$ 由所有同类样本特征取平均得到，k 表示类别个数。$\boldsymbol{d}_{\mathrm{intra}}$ 是融合后的特征 $\boldsymbol{F}$ 到本类中心的欧式距离，$\boldsymbol{d}_{\mathrm{inter}}$ 是融合后的特征表示 $\boldsymbol{F}$ 到其他类中心的欧式距离，具体如式(6)~式(7)所示：

$$\boldsymbol{d}_{\mathrm{intra}}=\left\|\boldsymbol{F}_K-\boldsymbol{C}_K\right\|_2\tag{6}$$

$$\boldsymbol{d}_{\mathrm{inter}}=\left\|\boldsymbol{F}_K-\boldsymbol{C}_{\bar{K}}\right\|_2\tag{7}$$

式中：$K,\bar{K}\in(1,2,3,4)$ 且 $K\neq\bar{K}$。

$\text{Loss}_{\text{Cluster}}$ 具体如式(8)所示：

$$\text{loss}_{\text{Cluster}}=-\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{d}_{\text{intra}}^{j}+\boldsymbol{d}_{\cos}^{j}) \tag{8}$$

式中：$\boldsymbol{d}_{\cos}$ 是 $\boldsymbol{d}_{\text{intra}}$ 与 $\boldsymbol{d}_{\text{inter}}$ 的余弦相似度。

边界检测辅助任务：实体边界信息指的是实体词组在文本序列中开始到结束的位置信息。多模态命名实体识别任务需要同时识别出实体词组的边界信息和实体类别信息。提升模型对实体词组的边界识别能力可以一定程度上促进命名实体的识别效果。因此，本文引入边界检测辅助任务。定义 $Z=(\text{z}_1,\text{z}_2,...,\text{z}_n)$ 为边界检测任务的标签。$\boldsymbol{X}$ 经过一层标准的 Transformer 层学习上下文信息，得到新表示 $\boldsymbol{H}=(\boldsymbol{h}_0,\boldsymbol{h}_1,...,\boldsymbol{h}_{n+1})$，其中 $\boldsymbol{h}_i\in\mathbb{R}^d$。随后，本文利用一层 CRF 来输出预测。得到预测序列标签 z′ 的概率及边界检测辅助任务的损失函数如式(9)~(10)所示：

$$\boldsymbol{P}(\text{z}'\mid\boldsymbol{H})=\frac{\exp(\text{score}(\boldsymbol{H},\text{z}'))}{\sum_{\text{z}^*\in Z^*}\exp(\text{score}(\boldsymbol{H},\text{z}^*))} \tag{9}$$

$$\text{loss}_{\text{EBD}}=-\frac{1}{N}\sum_{j=1}^{N}\log_{\text{a}}\boldsymbol{P}\left(\text{z}_j'\mid\boldsymbol{H}_j\right) \tag{10}$$

式中：Z^* 是所有可能标签序列合集，每种可能序列得分 $\text{score}\left(\boldsymbol{H},\text{z}'\right)$ 由转移得分和发射得分构成。

多模态命名实体识别：本文采用 CRF 结构来学习标签之间的依赖关系，来获得最优的预测序列。将融合后的特征 $\boldsymbol{F}$ 作为 CRF 层的输入，得到预测序列标签 y′ 的概率及多模态命名实体识别任务的损失函数如式(11)~ (12)所示：

$$P(\text{y}'\mid\boldsymbol{F})=\frac{\exp(\text{score}(\boldsymbol{F},\text{y}'))}{\sum_{\text{y}^*\in\text{Y}^*}\exp\left(\text{score}\left(\boldsymbol{F},\text{y}^*\right)\right)} \tag{11}$$

$$\text{loss}_{\text{MNER}}=-\frac{1}{N}\sum_{j=1}^{N}\log_{\text{a}}\boldsymbol{P}\left(\text{y}_j'\mid\boldsymbol{F}_j\right) \tag{12}$$

式中：Y^* 是所有可能标签序列合集，每种可能序列得分 $\text{score}\left(\boldsymbol{F},\text{y}'\right)$ 由转移得分和发射得分构成。

1.2 模型训练

模型训练过程中，主任务损失结合辅助任务损失共同来优化网络参数，损失函数如式(13)所示：

$$\text{Loss}=\text{loss}_{\text{MNER}}+\alpha\cdot\text{loss}_{\text{EDB}}+\beta\cdot\text{loss}_{\text{Fusion}}+\gamma\cdot\text{loss}_{\text{Cluster}} \tag{13}$$

式中：$\text{Loss}_{\text{MNER}}$ 是多模态命名实体识别任务损失，Loss_{EDB} 是边界检测辅助任务损失，$\text{Loss}_{\text{Fusion}}$ 是对比融合辅助任务损失，$\text{Loss}_{\text{Cluster}}$ 是实体聚类辅助任务损失。α，β，γ 分别为边界检测辅助任务、对比融合辅助任务、实体聚类辅助任务的系数。

2 实验结果与分析

2.1 数据集和评价指标

Twitter-2017[4]是多模态命名实体识别任务中经典数据集。Twitter-2017 来源于 Twitter 和 Snapchat，包含“Person”、“Location”、“Organization”、“Misc”4 类实体。其数据集的划分方式如表 1 所示。

表 1 Twitter-2017 划分信息

Tab.1 Divided information of Twitter-2017

数据集名称	训练集	验证集	测试集
Twitter-2017	3373	723	723

本文采用精准率(Precision，P)、召回率(Recall，R)和 F1 值评估多模态命名实体识别模型的有效性。

2.2 实验设置

据前人已有研究的参数设置，最大句子长度设置为 128，Batch-Size 为 32。文本嵌入部分使用 Bert-base-cased 预训练模型，图片嵌入使用 ResNet-152 预训练模型。其余参数详见表 2：

表 2 参数设置信息

Tab.2 Parameters setting

参数名称	α	β	γ	λ	t
Twitter-2017 参数值	0.4	0.5	0.01	0.75	0.1

2.3 实验结果分析

为了验证 MLMNER 模型的有效性，本文对比了经典的文本模态命名实体识别基线模型，以及多模态命名实体识别基线模型。

文本模态基线：CNN-BiLSTM-CRF[23]、BiLSTM-CRF[24]、HBiLSTM-CRF[25]以 BiLSTM 为模型主体结构，后续接 CRF 层学习文本序列转移概率。BERT-CRF 以 BERT 为模型主体结构，后续接 CRF 层学习文本序列转移概率。

图文模态基线：GVATT-HBiLSTM-CRF[4]、AdaCAN-CNN-BiLSTM-CRF[3] 利用 VGG-16 或 ResNet 抽取图片特征利用 BiLSTM 学习文本特征。UMT-BERT-CRF[6]利用 ResNet 抽取图片特征，使用 BERT 获取文本特征，后续接 CRF 层学习文本序列转移概率。

本文提出的模型与上述基线模型在 Twitter-2017 数据集上分别进行实验对比，实验结果如表 3 所示。

表 3　Twitter-2017 实验结果

Tab.3　Experimental results on the Twitter-2017

数据模态	方法名称	四类实体综合结果/%		
		P	R	F1
文本模态	BiLSTM-CRF	79.42	73.43	76.31
	CNN-BiLSTM-CRF	80.00	78.76	79.37
	HBiLSTM-CRF	82.69	78.16	80.37
	BERT-CRF	83.32	83.57	83.44
图文模态	GVAT-HBiLSTM-CRF	83.41	80.38	81.87
	AdaCAN-CNN-BiLSTM-CRF	84.16	80.24	82.15
	UMT-BERT-CRF	85.28	85.34	85.31
	MLMNER（本文）	**85.46**	**85.71**	**85.59**

由表 3 可知：

1）多任务学习的有效性。MLMNER 较 GVAT-HBiLSTM-CRF，在四类实体综合结果 F1 值上提升 3.72%，较 AdaCAN-CNN-BiLSTM-CRF，在四类实体综合结果的 F1 值上提升 3.44%。相较于以上两个单任务模型，本文方法有较大的提升。分析可知，本文方法中的辅助任务在模型学习过程中帮助主模型学习到丰富的实体知识，因此模型实验结果优于单任务模型，验证了多任务学习的有效性。

2)对比融合辅助任务和实体聚类辅助任务的有效性。MLMNER 较 UMT-BERT-CRF，在四类实体综合结果的精准率 P 上提升 0.18%，在召回率 R 上提升 0.37%，在 F1 值上提升 0.28%。分析可知，本文方法特有的对比融合辅助任务和实体聚类辅助任务在一定程度上促进了图文模态的融合，并提升了模型对易混淆实体的识别能力，验证了实体边界检测辅助任务和对比融合辅助任务的有效性。

2.4 消融实验

本节设计消融实验来验证两个辅助任务对实验结果的影响。本文方法（MLMNER）为基准模型，观察移除辅助任务（使用“No”表示移除该任务）后实验结果的变化。实验结果如表 4 所示。

表 4　Twitter-2017 消融实验结果

Tab.4　Ablation results on the Twitter-2017

消融方法	四类实体综合结果/%		
	P	R	F1
MLMNER（本文）	**85.46**	**85.71**	**85.59**
No 实体聚类	85.07	84.75	84.91(↓0.68)
No 对比融合	84.76	84.38	84.57(↓1.02)
No 实体聚类 & No 对比融合	83.19	84.23	83.71(↓1.88)

分析表 4 可知：

1）实体聚类辅助任务的有效性。MLMNER 去除实体聚类辅助任务时，四类实体综合 F1 值下降了 0.68%。说明了实体聚类辅助任务可帮助模型区分易混淆实体，提升模型实体识别能力。

2）对比融合辅助任务的有效性。MLMNER 去除对比融合辅助任务时，四类实体综合 F1 值下降了 1.02%。说明对比融合辅助任务可促进图文模态融合，提供更好的融合特征，提升模型实体识别能力。

3）多任务学习的有效性。MLMNER 去除实体聚类辅助任务与对比融合辅助任务时，四类实体综合 F1 值下降了 1.88%。说明多任务学习通过参数共享的方式帮助主任务学习相关知识，提升模型的实体识别能力。

3 结束语

本文提出一种基于多任务学习的多模态命名实体识别方法。首先，使用 Bert 和 ResNet 抽取原始文本和图片的特征，并利用跨模态 Transformer 结构融合多模态信息；随后，在主任务基础上，增加对比融合辅助任务促进图文模态信息融合；增加实体聚类辅助任务来学习实体类别之间的差异，提升模型对不同实体的区分能力。最后，通过实验证明了本文方法的有效性。虽然所提方法在当前数据集上实体识别能力有一定提升，但实验结果仍有较大的提升空间。随着多模态预训练语言模型的发展，多模态预训练语言模型逐渐被应用于自然语言处理领域。因此，下一步将探索多模态预训练语言模型在多模态命名实体识别任务中的应用。

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, 2022, 34(1): 50-70.

[2] MOON S, NEVES L, CARVALHO V. Multimodal named entity recognition for short social media posts[C]//Proceedings of the NAACL HLT 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies-Proceedings of the Conference. New Orleans, Louisiana, USA: Association for Computational Linguistics, 2018: 852-860.

[3] ZHANG Q, FU J, LIU X, et al. Adaptive co-attention network for named entity recognition in tweets[C]//Proceedings of the 32th AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA: AAAI Press, 2018: 5674-5681.

[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. Melbourne, Australia: Association for Computational Linguistics, 2018: 1990-1999.

[5] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE

conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE Computer Society, 2016: 770-778.

[6] 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. Online: Association for Computational Linguistics, 2020: 3342-3352.

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the NAACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, MN, USA: Association for Computational Linguistics, 2019: 4171-4186.

[8] KENDALL A, GAL Y, CIPOLLA R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: Computer Vision Foundation, 2018: 7482-7491.

[9] CLARK K, LUONG M-T, KHANDELWAL U, et al. BAM! Born-Again Multi-Task Networks for Natural Language Understanding[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 5931-5937.

[10] LI Y, CARAGEA C. Multi-Task Stance Detection with Sentiment and Stance Lexicons[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019: 6298-6304.

[11] WANG X, LYU J, DONG L, et al. Multitask learning for biomedical named entity recognition with cross-sharing structure[J]. Bioinform, 2019, 20(1): 427:1-427:13.

[12] ZHANG Y, YANG Q. An overview of multi-task learning[J]. National Science Review, 2018, 5(1): 30-43.

[13] CHEN Z, BADRINARAYANAN V, LEE C-Y, et al. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks[C]//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018: 794-803.

[14] YANG Y, HOSPEDALES T M. Deep Multi-task Representation Learning: A Tensor Factorisation Approach[C]//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: Conference Track Proceedings. OpenReview.net, 2017.

[15] REI M. Semi-supervised multitask learning for sequence labeling[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. Vancouver, Canada: Association for Computational Linguistics, 2017: 2121-2130.

[16] LIN Y, YANG S, STOYANOV V, et al. A multi-lingual multi-task architecture for low-resource sequence labeling[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics, 2018: 799-809.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems. Long Beach, CA, USA: MIT Press, 2017: 5998-6008.

[18] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the 37th International Conference on Machine Learning International conference on machine learning. Virtual Event: PMLR, 2020: 1597-1607.

[19] FANG H, WANG S, ZHOU M, et al. Cert: Contrastive self-supervised learning for language understanding[J]. arXiv preprint arXiv:2005.12766, 2020.

[20] GIORGI J, NITSKI O, WANG B, et al. DeCLUTR: Deep contrastive learning for unsupervised textual representations[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Virtual Event: Association for Computational Linguistics, 2021: 879-895.

[21] WU Z, WANG S, GU J, et al. Clear: Contrastive learning for sentence representation[J]. arXiv preprint arXiv:2012.15466, 2020.

[22] YOU Y, CHEN T, SHEN Y, et al. Graph Contrastive Learning Automated[J]. arXiv preprint arXiv:2106.07594, 2021.

[23] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: The Association for Computer Linguistics, 2016: 1064-1074.

[24] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[25] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego California, USA: The Association for Computational Linguistics, 2016: 260-270.