PAPER

Multimodal Named Entity Recognition with Bottleneck Fusion and Contrastive Learning

Peng WANG[†a)], ***Member***, **Xiaohang CHEN**[†b)], **Ziyu SHANG**[†c)], ***and*** **Wenjun KE**[†,††d)], ***Nonmembers***

SUMMARY Multimodal named entity recognition (MNER) is the task of recognizing named entities in multimodal context. Existing methods focus on utilizing co-attention mechanism to discover the relationships between multiple modalities. However, they still have two deficiencies: First, current methods fail to fuse the multimodal representations in a fine-grained way, which may bring noise of visual modalities. Second, current methods ignore bridging the semantic gap between heterogeneous modalities. To solve the above issues, we propose a novel MNER method with bottleneck fusion and contrastive learning (BFCL). Specifically, we first incorporate the transformer-based bottleneck fusion mechanism, subsequently, information between different modalities can only be exchanged through several bottleneck tokens, thus reducing the noise propagation. Then we propose two decoupled image-text contrastive losses to align the unimodal representations, making the representations of semantically similar modalities closer, while the representations of semantically different modalities farther away. Experimental results demonstrate that our method is competitive to the state-of-the-art models, and achieves 74.54% and 85.70% F1-scores on Twitter-2015 and Twitter-2017 datasets, respectively.

key words: *multimodal named entity recognition, contrastive learning, multimodal fusion, co-attention*

1. Introduction

Named entity recognition (NER) aims to detect the span of entities from a chunk of mention and classify them into pre-defined types, such as person (PER), location (LOC) and organization (ORG) [1]. It is a fundamental task in many downstream applications, such as event detection [2], knowledge graph construction [3] and entity linking [4]. With the rapid development of social media, massive users generate data in the form of text combined with images and videos. Large-scale data mining on these abundant multimodal data is challenging because the text in which cannot always provide enough context for traditional text-based NER methods to determine the types of entities. As shown in Fig. 1, if only according to the text *"Kolo jacky loves the sun"*, it is difficult for existing text-based NER methods to judge whether the type of entity *"Kolo"* is person or dog. Whereas, the related image clearly indicates that the entity *"Kolo"* is a dog. It can be seen that without multimodal information, many valuable content would be lost and entities cannot be correctly recognized, especially when the text is relatively short and coarse.

Text	Labels
Kolo	B-MISC
jacky	I-MISC
loves	O
the	O
sun.	O

Fig. 1 An example of multimodal named entity recognition. We use the BIO2 tagging schema [5] to tag the labels of the text, where B-MISC represents the start position of a MISC entity, I-MISC represents the inter position of a MISC entity and O represents non-entity.

To improve the text-based NER methods, multimodal named entity recognition (MNER) has attracted much attention recently, which aims at recognizing named entities from multimodal data. How to discover the auxiliary multimodal clues to detect the correct type of entities is one of the core issues of MNER. To this end, most existing works rely on co-attention mechanism to mine the relationships between modalities. Moreover, they also utilize gated multimodal fusion to integrate multiple unimodal representations into a joint multimodal representation. While previous works have shown success of mining multimodal relationships to facilitate MNER, they still suffer from two weaknesses:

- Firstly, most current methods ignore designing a fine-grained multimodal fusion mechanism for the MNER task. They use gated fusion to simply weight the whole visual representations and fuse them with textual representations as the classification features of entities. However, not all entities are relevant to visual representations, and the introduction of visual representations for these irrelevant entities would bring noise and affect the type features of entities.
- Secondly, current methods ignore bridging the semantic gap between modalities. Textual and visual representations are heterogeneous in that they come from different encoders and contain modality-specific semantic information. Heterogeneous representations prevent the MNER models from aligning the intuitively similar text locations and image regions. For example, in Fig. 1, the word *Kolo* should have a higher similarity

Manuscript received July 6, 2022.
Manuscript revised November 21, 2022.
Manuscript publicized January 18, 2023.

[†]The authors are with the School of Computer Science and Engineering, Southeast University, Nanjing, China.
[††]The author is with the Beijing Institute of Computer Technology and Application, Beijing, China.
a) E-mail: pwang@seu.edu.cn
b) E-mail: cxhang@seu.edu.cn
c) E-mail: ziyus1999@seu.edu.cn
d) E-mail: kewenjun2191@163.com

DOI: 10.1587/transinf.2022EDP7116

to the regions in the image that are related to the object dog than to other regions. However, due to the heterogeneous representations, when calculating the similarity score, the similarity between the textual representations of *Kolo* and the visual representations of dog may be lower than the similarity of other regions.

To address the above issues, we propose a novel MNER method with bottleneck fusion and contrastive learning (BFCL). Specifically, to fuse the multimodal representations in a fine-grained way, we incorporate the transformer-based bottleneck fusion mechanism by which information between different modalities can only be exchanged through several trainable bottleneck tokens, thus limiting the spread of noise. In a bottleneck fusion layer, we first concatenate the textual representations with bottleneck tokens and input them to transformer to capture the valuable information of textual modality. Then we concatenate the visual representations with bottleneck tokens and input them to another transformer to capture the valuable information of visual modality. Bottleneck tokens are updated twice, first with textual representations, and then with visual representations. By this way, textual and visual modalities can only interact with the shared bottleneck tokens to learn the beneficial features of other modalities. To bridge the semantic gap between modalities, we propose two decoupled image-text contrastive losses to align the unimodal representations of text and image, making the representations of semantically similar modalities closer, while the representations of semantically different modalities farther away.

The main contributions of this paper can be summarized as follows:

- We focus on designing a fine-grained multimodal fusion mechanism for MNER, introducing a transformer-based bottleneck fusion mechanism to fuse textual and visual representations in a fine-grained way.
- We propose two decoupled image-text contrastive losses to bridge the gap of heterogeneous textual and visual modalities and align the relevant unimodal representations.
- We perform comprehensive experiments and sensitivity analysis on two social media benchmark datasets. Results show that BFCL outperforms the strong state-of-the-art models and achieves 74.54% and 85.70% F1-scores on Twitter-2015 and Twitter-2017 datasets, respectively.

The rest of this paper is organized as follows. We will analyze related work in Sect. 2. Section 3 details the proposed method. Section 4 reports the experiments. Section 5 gives our conclusion.

2. Related Work

MNER is not only an extension of traditional NER task, but also involves special techniques: (1) multimodal co-attention and (2) multimodal fusion. In this section, we first briefly review the evolution of text-based NER models. Then we introduce existing MNER models based on multimodal co-attention mechanism. Finally, we summarize three multimodal fusion strategies.

2.1 Named Entity Recognition (NER)

As a crucial task of information extraction, NER has been widely studied in the past two decades. Typically, the NER task can be modeled as a sequence labeling problem. Various sequence labeling models have achieved good performance, including probabilistic graph models such as conditional random fields (CRF) [6], and deep neural networks like recurrent neural networks (RNN) [7] and convolutional neural networks (CNN) [8]. Recently, pre-trained language models (e.g., BERT [9]) demonstrate powerful feature extraction capability and produce great improvement in many downstream natural language processing tasks, especially NER. These models have achieved great performance in the textual modality. However, they still suffer from understanding multimodal data like social media posts where the text is typically informal and ambiguous.

2.2 Multimodal Named Entity Recognition

To improve the text-based NER models and utilize multimodal data, MNER models have been proposed. The goal of MNER is to mine the associations between multiple modalities and extract valuable information to make up for the lack of semantic context in a single modality. Existing MNER methods mainly make use of co-attention mechanism to combine textual and other modalities representations. Moon et al. [10] proposed a generic modality-attention module to attenuate irrelevant modalities and amplify the most informative ones. Zhang et al. [11] designed an adaptive co-attentive mechanism to combine information from textual and visual modalities and filter the inter-modal noise through a gated fusion mechanism. Wu et al. [12] first used Mask RCNN [13] to identify visual objects and convert them into the embedding vector space, and then designed a dense co-attention layer to model the guided attention between objects and entities. Yu et al. proposed a transformer-based [14] multimodal co-attention to obtain both image-aware word representations and word-aware visual representations. Xu et al. [15] proposed a general matching and alignment framework for MNER in social media posts.

These works have greatly improved the performance of MNER, and they usually use gated fusion mechanism to obtain unimodal representations. That may bring much noise because not all textual mentions are relevant to visual information. There are two reasons why an appropriate multimodal fusion mechanism is important to the MNER task. First, combining complementary information between modalities is beneficial to enrich the context of entities. Second, when one modality data is missing, it can use data of other modalities to identify entities. To this end, we use bottleneck fusion mechanism to force information between dif-

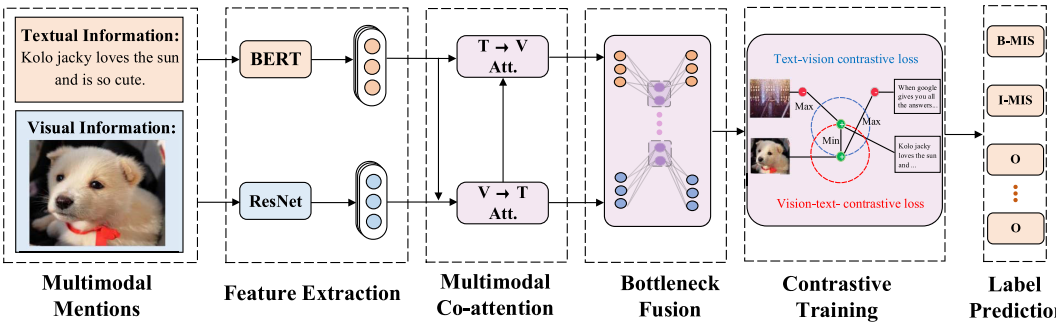


Fig. 2 Overview of the multimodal named entity recognition model BFCL.

ferent modalities to pass through a small number of bottleneck tokens. Instead of directly interacting with each other, multiple unimodal representations collate and condense the most relevant information in each modality and only share the necessary information through the bottleneck tokens.

2.3 Mutimodal Fusion

Multimodal fusion considers the process of integrating information of different modalities as joint representations learning [16]. According to the levels of fusion, there are three multimodal fusion strategies: early fusion, late fusion and middle fusion. Early fusion directly concatenates the original features of each modality, which is simple but ignores the heterogeneity among modalities. By contrary, late fusion first learns unimodal features from different modalities and then fuses learned features into a multimodal semantic representation to avoid the problem of heterogeneous features fusion [17], [18]. The disadvantages of late fusion schemes are its expensiveness in terms of the learning effort and the potential loss of correlation in mixed feature space. Compared with the early fusion and late fusion strategies, middle fusion is a more flexible and general fusion strategy [19]. Specifically, each modal is transformed from a separate neural layer as an input and in some subsequent hidden layers, multiple modal representations are mapped into the joint space. Therefore, the middle fusion is able to capture interactions among modalities by learning intermediate features and adding fully connected layers to continue training the concatenated features [20].

2.4 Contrastive Learning

Contrastive learning has been an effective training paradigm by pulling closer positive samples and pushing apart negative samples in the feature space using contrastive loss (e.g., InfoNCE loss [21]), which brings significant improvement for downstream tasks [21]–[24]. Specifically, contrastive learning was first introduced in computer vision to train visual representations [22], [23]. Subsequently, contrastive learning was utilized in natural language processing to leverage abundant textual resources to learn embeddings better [25]. Therefore, recent NER methods also try to utilize contrastive learning to enhance the performance of models. CONTaiNER [26] optimizes the inter-token distribution distance instead of class-specific attributes for Few-Shot NER. ConCNER [27] designs two contrastive objectives for cross-language NER at different grammatical levels. Different from these methods, in this paper, we leverage contrastive learning to enhance multimodal NER and propose two decoupled image-text contrastive losses to bridge the gap between heterogeneous textual and visual modalities.

3. Methodology

In this section, we first formally define the MNER task. Then we give a brief overview of the proposed BFCL model. Subsequently, we discuss the visual feature extraction and textual feature extraction. Finally, We elaborate the details of each module in BFCL.

3.1 Problem Formulation

In our work, we mainly study the MNER task containing textual and visual modalities. Given a text S and its associated image V as input, the aim of MNER is to recognize a set of named entities from S and classify each recognized named entity into one of the pre-defined types. As most existing work on MNER, we formulate the task as a sequence labeling problem. Let $S = (w_1, w_2, \ldots, w_n)$ denote a sequence of input words, and $Y = (y_1, y_2, \ldots, y_n)$ be the corresponding label sequence, where $y_i \in \mathcal{Y}$ and $\mathcal{Y}$ is the pre-defined label set with the BIO2 tagging schema [5]. For example, in Fig 1, we use B-MISC to tag the start position of a MISC entity, I-MISC to tag the inter position of a MISC entity and O to tag non-entity. Notice that although this paper discusses two modalities, the proposed model is able to be extended to handle more modalities.

3.2 Overview of BFCL

The structure of the BFCL model is shown in Fig. 2. BFCL includes five steps: (1) multimodal feature extraction, (2) multimodal co-attention, (3) bottleneck fusion, (4) contrastive training and (5) label prediction. Firstly, we respectively extract visual features by ResNet [28] and embed token textual features with BERT. Afterwards, multimodal co-attention mechanism alternately focuses on multimodal

information, constructing visual-guided textual representations and textual-guided visual representations. Bottleneck fusion is then used to adaptively integrate the above co-attended textual and visual representations. Furthermore, we design vision-text contrastive loss and text-vision contrastive loss during training to improve the quality of multimodal representations. Finally, we input the fused textual representations to a CRF layer to predict the label of each token.

3.3 Multimodal Feature Extraction

Because different modalities have different semantic information, we first extract features from textual and visual modalities.

3.3.1 Textual Features

Textual features are extracted with BERT [9] and fine-tuned in training. Formally, let $S = [s_1, s_2, \ldots, s_n]$ be the decomposed sentence input tokens, where n denotes the length of the input tokens, s_1 and s_n denote the two inserted special tokens "[CLS]" and "[SEP]" that denote the start and end position of a sentence. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ represent the vector representation of the sentence S, where $\mathbf{x}_i$ is the vector representation of i-th token, which is the sum of the token, segment and position embedding. $\mathbf{X}$ is then input to the BERT encoder to get the textual representations $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n]$, where $\mathbf{c}_i \in \mathbb{R}^{d_t}$ is the context-aware representation of the i-th token extracted by BERT, and d_t is the dimension of BERT output vector.

The dimensions of multimodal representations output by different modal encoders are inconsistent. In order to facilitate subsequent multimodal correlation discovery and fusion, we unify the dimensions of multimodal representation by using a fully connected layer to map the textual representations to multimodal representations space as follows:

$$\mathbf{T} = \mathrm{ReLU}\left(W_t\mathbf{C} + b_t\right) \tag{1}$$

where $W_t \in \mathbb{R}^{d_t \times d}$ is the weights, $b_t \in \mathbb{R}^d$ is the bias term, $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n]$ is the textual representations transformed by the fully connected layer.

3.3.2 Visual Features

Similar to existing MNER works, visual representations are extracted through the pretrained ResNet [28]. Formally, given an image, we first resize it to 224×224 pixels and then input it to ResNet. ResNet splits the input image into $7 \times 7 = 49$ visual blocks with the same size and represents each block with a 2048-dimensional vector. Let $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{49})$ be the visual representations output by ResNet, where $\mathbf{u}_i \in \mathbb{R}^{2048}$ is the representations of the i-th region of the image. In order to facilitate the fusion with textual features, a fully connected layer is used to map visual representations into multimodal representations space:

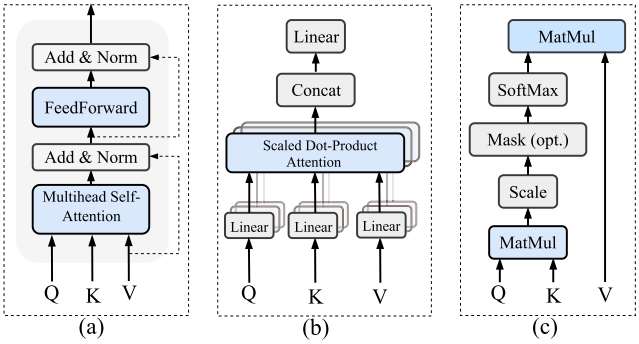


Fig. 3 Multimodal co-attention: (a) Cross-Modal Transformer; (b) Multihead Self-Attention; (c) Scaled Dot-Product Attention.

$$\mathbf{V} = \mathrm{ReLU}\left(W_v\mathbf{U} + b_v\right) \tag{2}$$

where $W_v \in \mathbb{R}^{2048 \times d}$ is the weights, $b_v \in \mathbb{R}^d$ is the bias term and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{49}]$ is the transformed visual representations.

3.4 Multimodal Co-Attention

Multimodal co-attention (MCA) is designed to capture the fine-grained correlations between modalities by calculating the attention weights between all textual positions and all visual regions. MCA is composed of two attention mechanisms: textual-guided visual attention and visual-guided textual attention, which alternately guide each other to gather useful information from different modalities. Both of them are built upon on the Cross-Modal Transformer (CMT) layer [29].

As shown in Fig. 3 (a), the most important component of CMT is multi-head self-attention (MH-SA) mechanism. It can be seen from Fig. 3 (b) that MH-SA is stacked by scaled dot-product attention. Figure 3 (c) is the structure of scaled dot-product attention which takes query (Q), key (K) and value (V) as input, and then output the weighted sum of value. In visual-guided textual attention, we input image representations $\mathbf{V} \in \mathbb{R}^{49 \times d}$ as query, text representations $\mathbf{T} \in \mathbb{R}^{n \times d}$ as key and text representations $\mathbf{T} \in \mathbb{R}^{n \times d}$ as value. Then the scaled dot product attention can be calculated as follows:

$$\mathrm{A}(\mathbf{V}, \mathbf{T}, \mathbf{T}) = \mathrm{softmax}\left(\frac{\mathbf{V}\mathbf{T}^\top}{\sqrt{d}}\right)\mathbf{T} \tag{3}$$

Given visual representations $\mathbf{V}$ and textual representations $\mathbf{T}$, $\mathbf{V}\mathbf{T}^\top \in \mathbb{R}^{49 \times n}$ represents the dot-product similarity matrix of vision and text, and $\mathbf{V}\mathbf{T}^\top_{ij}$ represents the similarity between the j-th textual position and the i-th visual region. By calculating the local similarities, our model can mine the fine-grained relationships between textual and visual information. Multi-head self-attention projects query, key and value into different representations spaces by linear layer, each space can be called a head, and each head calculates the scaled dot-product attention in parallel, allowing the model

to mine the correlations between multimodal representations from different aspects in a fine-grained manner. Let h be the number of heads, in order to improve the computational efficiency, the dimension of each head is set to d/h and the output of MH-SA is obtained by concatenating the h results as follows:

$$\text{MH-SA}(\mathbf{V},\mathbf{T},\mathbf{T}) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \quad (4)$$

$$\text{head}_i = \text{A}(\mathbf{V}W_i^Q, \mathbf{T}W_i^K, \mathbf{T}W_i^V) \quad (5)$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_h}$, $W^O \in \mathbb{R}^{d\times d}$ are projection matrices for the i-th head.

In addition, CMT also employs a feed-forward layer, residual connection and layer normalization to facilitate optimization. In the visual-guided textual attention, the whole operation of CMT can be calculated as follows:

$$\tilde{\mathbf{T}}_\mathbf{v} = \text{LN}(\mathbf{V} + \text{MH-SA}(\mathbf{V},\mathbf{T},\mathbf{T})) \quad (6)$$

$$\mathbf{T}_\mathbf{v} = \text{LN}(\tilde{\mathbf{T}}_\mathbf{v} + \text{FFN}(\tilde{\mathbf{T}}_\mathbf{v})) \quad (7)$$

where FFN is the feed-forward network, LN is the layer normalization, $\mathbf{T}_\mathbf{v} \in \mathbb{R}^{49\times d}$ is the visual-guided text representations. Then we input $\mathbf{T}_\mathbf{v}$ to the CMT layer again to change its dimension:

$$\mathbf{T}_\mathbf{v} = \text{CMT}(\mathbf{T},\mathbf{T}_\mathbf{v},\mathbf{T}_\mathbf{v}) \quad (8)$$

Similarly, the textual-guided visual representations $\mathbf{V}_\mathbf{t} \in \mathbb{R}^{n\times d}$ can be obtained as follows:

$$\mathbf{V}_\mathbf{t} = \text{CMT}(\mathbf{T},\mathbf{V},\mathbf{V}) \quad (9)$$

3.5 Bottleneck Fusion

Bottleneck fusion is responsible for fusing the representations of multiple modalities into a joint representation, which includes specific features within each modal, correlations between modalities, and complementary information. The traditional multimodal fusion methods can be divided into three types according to the fusion timing: early fusion, late fusion and middle fusion. Early fusion fuses the original data features of different modalities and then inputs them into the model. As depicted in Fig. 4 (a), late fusion aims to fuse the final representations of modalities. It can be seen from Fig. 4 (b) that middle fusion is to fuse the intermediate representations encoded by the respective encoders of each modal and then input it to the subsequent network.

Although the above three fusion methods have different fusion timings, they all directly interact with the representations of different modalities during fusion. However, most information of a modality is redundant and irrelevant for other modalities, and direct fusion will introduce a lot of noise. We use bottleneck fusion to solve this problem. Similar to the traditional multimodal fusion methods, it can be divided into bottleneck-early fusion, and bottleneck-middle fusion. The principle of bottleneck fusion is shown in Fig. 4 (c) and Fig. 4 (d). By setting a small number of fusion units (called bottlenecks), the information between modalities is not directly interacted but only passed through the bottleneck, requiring the model to organize and condense the most relevant information in each modality, sharing only the necessary information. Using the bottleneck fusion method can effectively limit the propagation of noise and improve the efficiency of fusion.

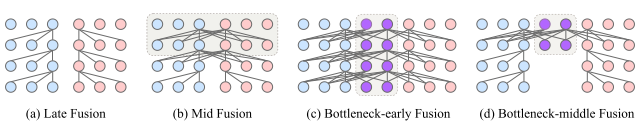


Fig. 4 Structure of three fusion mechanism. Circles with blue, pink and purple represent the textual representations, visual representations and bottleneck tokens, respectively.

We realize the bottleneck fusion by stacking L layers of transformers. Specifically, we first initialize m fusion bottlenecks $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m]$, where $\mathbf{b}_i \in \mathbb{R}^d$ is the i-th bottleneck. For the l-th layer bottleneck fusion, the textual and visual representations are updated by:

$$\left[\mathbf{T}_\mathbf{v}^{l+1}, \hat{\mathbf{B}}^{l+1}\right] = \text{Transformer}\left(\left[\mathbf{T}_\mathbf{v}^{l}, \mathbf{B}^{l}\right]; \theta_{\text{text}}\right) \quad (10)$$

$$\left[\mathbf{V}_\mathbf{t}^{l+1}, \mathbf{B}^{l+1}\right] = \text{Transformer}\left(\left[\mathbf{V}_\mathbf{t}^{l}, \hat{\mathbf{B}}^{l+1}\right]; \theta_{\text{vision}}\right) \quad (11)$$

where $[\cdot,\cdot]$ denotes the concatenation of two tokens, $\mathbf{T}_\mathbf{v}^{l+1} \in \mathbb{R}^{n\times d}$ is the textual representations after the l-th layer bottleneck fusion with the visual information, $\mathbf{V}_\mathbf{t}^{l+1} \in \mathbb{R}^{n\times d}$ is the visual representations after the l-th layer bottleneck fusion with the textual information. θ_{text} and θ_{vision} are the parameters contained in the textual and visual transformer, respectively. Here the textual representations $\mathbf{T}_\mathbf{v}^{l+1}$ and the visual representations $\mathbf{V}_\mathbf{t}^{l+1}$ can only exchange information through the bottlenecks $\mathbf{B}$ within the transformer layer, which is updated twice. Since the task is to extract entities from the text modal, we believe that the importance of textual representations is higher, so the text representations $\mathbf{T}_\mathbf{v}^{L} \in \mathbb{R}^{n\times d}$ is used as the final multimodal fusion representation: $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n]$.

3.6 Label Prediction

After obtaining the multimodal representations of each token of the text, and given the input sentence S and its associated image I, we then apply a CRF layer to predict the probability of the label sequence $\mathbf{y}$ as follows:

$$P(\boldsymbol{y} \mid S, I) = \frac{\exp\left(\sum_{i=1}^{n} E_{h_i,y_i} + \sum_{i=0}^{n} T_{y_i,y_{i+1}}\right)}{\sum_{y}\exp(\sum_{i=1}^{n} E_{h_i,y_i} + \sum_{i=0}^{n} T_{y_i,y_{i+1}})} \quad (12)$$

where E_{h_i,y_i} is the emission score of label y_i for the i-th token, $T_{y_i,y_{i+1}}$ is the transition score from label y_i to label y_{i+1}.

3.7 Contrastive Training

To bridge the gap between the heterogeneous modalities and improve the quality of the learned multimodal representations, we design two decoupled contrastive losses: text-vision contrastive loss and vision-text contrastive loss. Given a batch of text-vision pairs, for the i-th pair, (t_i, v_i) is regarded as a positive pair, while other pairs in the batch are

regarded as negative pairs. We aim to maximize the similarity of the positive pairs and minimize the similarity of the negative pairs by minimizing the two contrastive losses. For example in the contrastive training step of Fig. 2, we aim to fine-tune related representations of the image of the dog and the text about "Kolo jacky" to be close, while staying away from the unrelated representations.

Formally, the vision-text contrastive loss is defined as follows:

$$\mathcal{L}_{vt} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(v_i,t_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(v_i,t_j)/\tau} + e^{\text{sim}(v_i,t_i)/\tau}} \tag{13}$$

$$\text{sim}(v_i, t_i) = \frac{f(v_i)^\top g(t_i)}{\|f(v_i)\| \cdot \|g(t_i)\|} \tag{14}$$

where N is the number of examples, τ is the temperature parameter, $f(\cdot)$ and $g(\cdot)$ is two modality encoders, $\text{sim}(\cdot)$ is the similarity measure function. Similarly, the text-vision contrastive loss is defined as follows:

$$\mathcal{L}_{tv} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(t_i,v_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(t_i,v_j)/\tau} + e^{\text{sim}(t_i,v_i)/\tau}} \tag{15}$$

Let $D = \left\{S_j, I_j, \mathbf{y}_j\right\}_{j=1}^{N}$ denote the training examples. We use the negative log-likelihood loss as the main loss function to train the proposed MNER model as follows:

$$\mathcal{L}_{\text{MNER}} = -\frac{1}{|D|} \sum_{j=1}^{N} (\log P(\boldsymbol{y}_j \mid S_j, I_j)) \tag{16}$$

In summary, the total training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{MNER}} + \lambda_{vt}\mathcal{L}_{vt} + \lambda_{tv}\mathcal{L}_{tv} \tag{17}$$

where λ_{vt} and λ_{tv} are trade-off parameters in two contrastive losses.

4. Experiments

This section reports the experimental results. We first address the experimental settings, including the datasets, baseline methods, parameter settings, and evaluation metrics. Then, we compare BFCL with state-of-the-art MNER models on the datasets. Furthermore, we conduct the ablation study and detailed analysis of the components in BFCL. Finally, we perform detailed case study and limitation analysis.

4.1 Datasets, Baselines and Settings

4.1.1 Datasets

To the best of our knowledge, the only two public MNER datasets are Twitter-2015 [11] and Twitter-2017 [30], which are both collected from Twitter. Therefore, we conduct experiments on the two datasets to evaluate the performance of the proposed model. Each example in the two datasets consists of a piece of text and an associated image. Table 1 shows the statistics of Twitter-2015 and Twitter-2017, both of which contain four entity types: person (PER), location (LOC), organization (ORG), and miscellaneous (MISC). Twitter-2015 contains 8,257 samples and 12,800 entities, and Twitter-2017 contains 4,819 samples and 8,724 entities. It can be seen that the size of Twitter-2015 is about 1.7 times that of Twitter-2017. In addition, the four types of entities are unevenly distributed in both datasets. In Twitter-2015, person and location entities account for a larger proportion, about 69%. In Twitter-2017, person and organization entities account for a larger proportion, about 76%. The training, development, and test sets are splited by Zhang et al. [11] and Lu et al. [30].

Table 1 Summary statistics of the datasets.

Entity type	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	1,546	5,078	6,049	1,324	1,351
#Samples	4,000	1,000	3,257	3,373	723	723

4.1.2 Baselines

We compare BFCL with the below strong text-only and text+vision baseline models. The text-only NER models are listed as follows:

- BiLSTM-CRF [7], which is the classical NER model that combines BiLSTM [31] and CRF.
- CNN-BiLSTM-CRF [8], which improves BiLSTM-CRF by replacing the embedding of each token with the concatenation of its word embedding and CNN-based character-level word representations.
- BERT [9], which is the pre-trained language model stacked by multiple layers bidirectional transformer, followed a softmax layer for predictions.
- BERT-CRF, which uses BERT as the encoder and replace the softmax layer by CRF.
- BERT-BiLSTM-CRF, which is an extension on the basis of BiLSTM-CRF. It feeds the contextual representations of BERT to BiLSTM to obtain the hidden representations of tokens.

The text+vision MNER models are listed as follows:

- GVATT [30], which exploits visual attention and gate mechanism to mine latent information from the entire image to guide word representation learning based on hierarchical BiLSTM-CRF.
- AdaCAN [11], which designs an adaptive co-attention network to induce word-aware visual representations for each word.
- GVATT-BERT-CRF, the word embedding of which is replaced with the contextual representations output by

Table 2 Performance comparison on two MNER datasets. Baselines with † are retrieved from [32]. Baselines with ‡ are retrieved from the original papers. The rest baselines are reproduced according to corresponding papers. Results of all the models are the average of random three times.

Modality	Model	Twitter-2015							Twitter-2017						
		Single Type (F1)				Overall			Single Type (F1)				Overall		
		PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1
Text	BiLSTM-CRF† [7]	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31
	CNN-BiLSTM-CRF† [8]	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
	BERT [9]	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
	BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
	BERT-BiLSTM-CRF	84.32	79.31	61.66	37.53	71.03	73.57	72.27	90.29	84.55	80.97	64.85	83.20	84.68	83.93
Text+Vision	GVATT† [30]	82.66	77.21	55.06	35.25	73.96	67.90	70.80	89.34	78.53	79.12	62.21	83.41	80.38	81.87
	AdaCAN† [11]	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
	GVATT-BERT-CRF†	84.43	80.87	59.02	38.14	69.15	74.46	71.70	90.94	83.52	81.91	62.75	83.64	84.38	84.01
	AdaCAN-BERT-CRF†	85.28	80.64	59.39	38.88	69.87	74.59	72.15	90.20	82.97	82.67	64.83	85.13	83.20	84.10
	UMT† [32]	85.24	81.58	63.03	39.45	71.84	74.61	73.20	**91.56**	84.73	82.24	**70.10**	85.08	85.27	85.18
	OCSGA‡ [12]	84.68	79.95	56.64	39.47	**74.71**	71.21	72.92	91.19	82.72	81.40	67.54	83.45	85.49	84.46
	ATTR-MMKG-MNER‡ [33]	84.28	79.43	58.97	**41.47**	74.78	71.82	73.27	-	-	-	-	-	-	-
	BFCL	**85.60**	**81.77**	**63.81**	40.30	74.02	**75.07**	**74.54**	91.17	**86.43**	**83.97**	66.67	**85.99**	**85.42**	**85.70**

BERT.

- AdaCAN-BERT-CRF, which uses BERT as the encoder to replace the original CNN and BiLSTM.
- OCSGA [12], which first uses Mask RCNN [13] to identify and embed visual objects in the image. Then it designs a dense co-attention layer to model the correlations between visual objects and textual features and the internal connections of objects or entities
- UMT [32], which implements a transformer-based multimodal co-attention mechanism, and an auxiliary entity span detection module to alleviate the visual bias.
- ATTR-MMKG-MNER [33], which is a multimodal NER model that introduces both image attributes and image knowledge to help improve NER task.
- BFCL, which is the model we proposed in this paper.

4.1.3 Settings

For features extraction, we use the pre-trained $BERT_{base}$ as the textual encoder and fine-tuned during training, the maximum text length is set to 128. The image encoder is the pre-trained ResNet-152 and not fine-tuned during training. The dimension of the multimodal space is set to 768. For multimodal co-attention, we set 12 self-attention heads, 1 transformer layer. For bottleneck fusion, we choose the number of bottlenecks and bottleneck fusion layers via a small grid search over $[2, 4, 6, 8, 16, 32, 64, 128]$ and $[0, 2, 4, 6, 8, 10, 12]$, respectively. For training, we set the temperature parameter to 0.6 and use the dot production as the similarity measure function. The learning rate, the dropout rate, batch size are respectively set to 5e-5, 0.1, and 16. The above hyperparameters are obtained through a small grid search in the development set of all datasets, which has the best F1 score. All the experiments are conducted on NVIDIA GTX 2080 Ti GPUs with PyTorch 1.8.1 and Python 3.7.11.

We use F1 score as the evaluation metric. The calculation formula of F1 score is as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{18}$$

where precision represents the probability of the truly positive samples among all the predicted positive samples, recall represents the probability of the predicted positive samples among the truly positive samples.

4.2 Main Results

We conduct comparative experiments to evaluate the effectiveness of each model in the MNER task. Table 2 shows the metrics of F1 score of every single type and overall precision (P), recall (R) and F1 score of all models on two datasets.

Firstly, we observe that using BERT as the textual encoder can improve the performance of models. Models using BERT generally outperform other models, where the F1 scores of BERT on two datasets are 4.17% and 3.58% higher than CNN-BiLSTM-CRF, respectively. This shows that the textual encoder is of great significance to the MNER task. Through transfer learning, the output representations of BERT contain much general knowledge of natural language, which is more suitable for scenarios with insufficient text information and noisy data compared to CNN and LSTM.

Secondly, it is noticed that adding visual features can bring valuable information mostly. Models that introduce visual modal, such as GVATT and AdaCAN, are better than BiLSTM-CRF and CNN-BiLSTM-CRF that only use text modal data. Especially, there is a huge improvement in precision, which means that when the text information is insufficient, the introduction of visual information can better help identify the type of entities. However, we also note that GVATT-BERT-CRF performs worse than BERT-CRF in Twitter-2015, and better in Twitter-2017. The reason is probably that visual modal information and textual modal information in Twitter-2017 are more consistent, and there

is relatively less noise in the visual modal that is not described in the text modal. This demonstrates the importance of dealing with the noise brought about by other modal data.

Thirdly, it can be seen that transformer-based multimodal co-attention mechanism is more suitable for mining the associations between multimodal representations. The performance of UMT and BFCL which use the transformer-based co-attention, greatly surpass AdaCAN and OCSGA which use the basic co-attention mechanism on two datasets. This suggests that the multi-head self-attention mechanism of transformer can compute the correlation information between each position of the text and region of image in a more fine-grained way.

Finally, we find that BFCL achieves state-of-the-art performance on both datasets, indicating the effectiveness of our model. Specifically, BFCL surpasses the second best model UMT 1.34% on Twitter-2015 and 0.52% on Twitter-2017 in overall F1 score. The reasons why F1 score improvements in Twitter-2015 and Twitter-2017 datasets have large gaps are as follow: (1) There is some noise in the visual modal information in Twitter-2015, which is not described in the corresponding textual modal information. (2) The average text length of Twitter-2015 is shorter than Twitter-2017, making it more difficult to align relevant unimodal representations in noisy environments for previous methods. Moreover, experimental results demonstrates that the two modules and contrastive losses we proposed can benefit the model to better fuse textual representations and visual representations.

4.3 Ablation Experiments

We conduct ablation experiments to demonstrate the effectiveness of each component in BFCL with F1 scores (%) reported in Table 3.

Firstly, we remove the bottleneck fusion module and directly input the visual guided textual representations to CRF. The F1 scores of BFCL drop 1.10% on Twitter-2015 and 0.65% on Twitter-2017, which indicates that bottleneck fusion plays an essential role in our model.

Secondly, we do not add the text-vision and vision-text contrastive losses to the negative log-likelihood loss during training. The F1 scores of BFCL drop 0.88% on Twitter-2015 and 0.49% on Twitter-2017, which indicates that contrastive losses can help fine-tune the parameters of encoder and bridge the gap between heterogeneous multimodal representations.

Finally, we remove the multimodal co-attention module and input the output of BERT encoder and ResNet encoder to the bottleneck fusion module. The F1 scores of BFCL drop 0.59% on Twitter-2015 and 0.34% on Twitter-2017, which demonstrates the importance of mining the relationships between textual and visual modalities.

4.4 Modality Encoder and Fusion Mechanism

Few works have analyzed the effects of visual modality encoders to the MNER task, which are important for subsequent multimodal associations discovery and fusion. To this end, we report the performance of the BFCL model under different visual modality encoders in Table 4. On Twitter-2015, the F1 score of BFCL gradually increases, as the layers of ResNet increases from 50 to 152. But BFCL-ResNet-200 performs worse than BFCL-ResNet-152. On Twitter-2017, BFCL-ResNet-152 gets the best performance and BFCL-ResNet-50 outperforms BFCL-ResNet-101 and BFCL-ResNet-200. The results suggest that (1) the performance of the MNER model does not only depend on the textual encoder, but also is sensitive to the used image encoder; (2) it is important to choose a visual encoder with appropriate layers.

Furthermore, to study the effect of multimodal fusion mechanism to the MNER task, Table 5 shows the precision, recall and F1 score of BFCL using three fusion strategies, namely, attention fusion [34], gated fusion [32], and bottleneck fusion. Attention fusion is to use attention mechanism to calculate the importance of each modality, and perform weighted sum according to the importance to fuse multiple modalities. Gated fusion is to input the multimodal embeddings to a full connected layer to get a logit as the importance of each modality and perform weighted sum according to the importance to fuse multiple modalities. Bottleneck fusion is to use several shared bottlenecks to exchange information among modalities. For example, the bottlenecks and embeddings of modality A are first concatenated and

Table 3 Ablation experiments results of BFCL on two datasets. BF denotes the bottleneck fusion module. CL denotes the contrastive loss. MCA denotes the multimodal co-attention module.

Models	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
BFCL	74.02	75.07	74.54	85.99	85.42	85.70
- BF	69.22	74.59	73.44	85.27	84.83	85.05
- CL	68.30	74.61	73.66	85.57	84.85	85.21
- MCA	66.24	68.09	73.95	85.29	85.42	85.36

Table 4 Comparison of BFCL using different modality encoder.

Visual Encoder	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
ResNet-50	73.85	73.80	73.83	85.71	84.38	85.04
ResNet-101	73.29	74.57	73.93	84.49	84.01	84.45
ResNet-152	74.02	75.07	74.54	85.99	85.42	85.70
ResNet-200	73.61	74.49	74.05	86.05	83.57	84.79

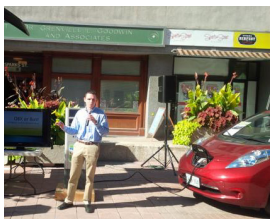
Table 5 Comparison of BFCL using different fusion mechanisms. Mech. denotes the fusion mechanism. AF denotes the attention fusion proposed by [34]. GF denotes the gated fusion used in [32]. BF denotes the bottleneck fusion.

Mech.	Twitter-2015			Twitter-2017		
	P	R	F1	P	R	F1
AF	69.22	74.59	73.44	85.27	84.83	85.05
GF	68.30	74.61	73.66	85.57	84.85	85.21
BF	74.02	75.07	74.54	85.99	85.42	85.70

(a) [Kolo] loves the sun and is so pretty, too.

(b) [Sydney] is at [Citrus College], his first gymnastics showcase !

(c) When [google] gives you all the answers to your homework.



(d) Thanks [Andrew] for the gre -at [Tesla] road trip presentation.

BERT-CRF:	1-PER	✗	1-LOC	✗	2-ORG	✗	1-ORG	✓	1-PER	✓	2-None	✗
UMT:	1-MISC	✓	1-PER	✓	2-ORG	✗	1-LOC	✗	1-PER	✓	2-PER	✗
BFCL:	1-MISC	✓	1-PER	✓	2-LOC	✓	1-ORG	✓	1-PER	✓	2-PER	✗

Fig. 5 Multimodal named entity recognition cases.

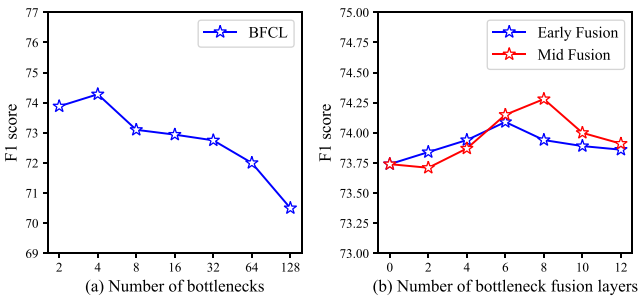


Fig. 6 Parameter sensitivity analysis experiment of the bottleneck fusion module. Mid Fusion denotesmiddle fusion.

input to a transformer to learn the important information of modality A and then the bottlenecks will be concatenated with modality B to another transformer to fuse the important information. On Twitter-2015, the F1 score of BFCL using the bottleneck fusion is about 0.9% higher than that of BFCL-GF using gated fusion, probably because the bottleneck fusion mechanism restricts the flow of information and only fuses the key information among multiple modalities representations. At the same time, this gap also highlights that multimodal fusion has a certain impact on the MNER task, and further proves that the bottleneck fusion mechanism is a more effective multimodal fusion mechanism.

4.5 Parameter Sensitivity Analysis

In order to more thoroughly understand the characteristics of the bottleneck fusion module and the impact of parameter changes, a parameter sensitivity analysis is carried out especially for this module.

We change the number of fusion bottlenecks and show the fluctuation of the F1 score of BFCL on Twitter-2015. Specifically, we set the number of fusion bottlenecks $|B| = [2, 4, 8, 16, 32, 64, 128]$, and fix the number of bottleneck fusion layers to 6. After training 30 epochs, the model with the best performance on the dev set is selected for testing. As shown in Fig. 6 (a), it is interesting that BFCL achieves the optimal result with only 4 bottlenecks. As the number of bottlenecks increases, the performance of BFCL first decreases steadily and decreases significantly at 64, indicating that if we set too many bottlenecks, it may make multimodal representations interact overly and much worthless information will be fused, which further illustrates the importance of limiting the flow of information between modalities.

We then investigate the effect of changing the number of bottleneck fusion layers to BFCL on Twitter-2015. Specifically, we change the number of layers under two different fusion strategies: (1) Early fusion, where the textual and visual representations processed by MCA module directly perform L layers bottleneck fusion; (2) Mid fusion, where the textual and visual representations processed by MCA module first pass through L/2 layers self-attention, and then perform L/2 layers bottleneck fusion. During the experiment, the number of bottlenecks is fixed 4. The change of F1 score of BFCL under the two strategies is shown in Fig. 6 (b). It can be seen that when the number of layers is more than 8, the model using the mid fusion strategy outperforms the model using the early fusion strategy. When the number of layers reaches 8, the mid fusion model achieves the best performance.

4.6 Case Study

To more intuitively demonstrate the capability of BFCL, we select four typical extraction cases as shown in Fig. 5 and give the prediction results of the BERT-CRF, UMT and BFCL.

Case (a) illustrates that the visual representations have valuable information that can help determine the entity's type. The text "Kolo loves the sun and is so pretty, too" has great ambiguity and lacks sufficient evidence to determine the type of entity "Kolo". Humans can only judge the entity "Kolo" as a person based on common sense. BERT-CRF, which only uses text information, misidentifies the entity "Kolo" as PER, while the multimodal models UMT and BFCL both correctly identify the entity "Kolo" as MISC, because they find the associations between text and image.

Case (b) illustrates that BFCL has stronger multimodal correlations mining ability. The first entity "Sydney" is the same English language as the place "Sydney", which BERT-

CRF incorrectly identifies as LOC. UMT and BFCL understand the semantic correspondence between text and image existence: a person performs gymnastics in a gym, correctly identifying "Sydney" as a PER. For the second entity "Citrus College", both BERT-CRF and UMT determine it to be ORG, while BFCL correctly recognizes "Citrus College" as Loc.

Case (c) illustrates that BFCL has a stronger ability to filter visual noise. The text expression means "I found the answer to my homework on Google", while the visual information depicts the scene of the concert, and there is no semantic matching between the two. So the visual information becomes noise, causing UMT to incorrectly determine "google" as PER, but BFCL has a certain noise filtering ability through bottleneck fusion and contrastive losses, and correctly identifies "google" as ORG.

Case (d) shows the misrecognition problem of BFCL. The text expresses the meaning of thanks to "Andrew" for his Tesla road trip demonstration, and there is a semantic correspondence with the person in the image standing next to the car with a microphone, so BFCL believes that text and visual information are highly related. As a result, the subject information in the image is excessively introduced, and "Tesla" is misidentified as PER.

5. Conclusion

In this paper, we focus on the rarely studied multimodal fusion method for multimodal named entity recognition and propose a novel MNER model BFCL, which improves state-of-the-art performance on social media posts. Specifically, we first extract the unimodal representations by textual and visual encoder and use the transformer-based multimodal co-attention to discover the relationships between modalities. Then we incorporate the bottleneck fusion to fuse the unimodal representations through some trainable shared tokens. Besides, to make the text representations and image representations more consistent, we propose two multimodal contrastive losses to help fine-tune the modality encoder. We conduct abundant experiments to show the great performance of BFCL and the importance of each module.

In addition, a limitation of our model is that it fails to alleviate the negative effects of visual information. In the future, we will focus on studying what information in the visual representations is helpful for MNER and filter out the potential noise from other modalities. Moreover, even this paper mainly discusses textual and visual modalities, our model is able to be extended to process more modalities.

References

[1] X. Li, J. Feng, Y. Meng, Q. Han, F. Wu, and J. Li, "A unified MRC framework for named entity recognition," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.5849–5859, 2020.

[2] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang, "Cross-media structured common space for multimedia event extraction," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.2557–2568, 2020.

[3] X. Fu, J. Zhang, H. Yu, J. Li, D. Chen, J. Yuan, and X. Wu, "A speech-to-knowledge-graph construction system," Proceedings of the 29th International Joint Conference on Artificial Intelligence, pp.5303–5305, 2020.

[4] P. Wang, J. Wu, and X. Chen, "Multimodal entity linking with gated hierarchical fusion and contrastive training," Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.938–948, 2022.

[5] E.F.T.K. Sang and J. Veenstra, "Representing text chunks," Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, pp.173–179, 1999.

[6] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," Proceedings of the 30th Conference on Computational Natural Language Learning, pp.147–155, 2009.

[7] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," CoRR, vol.abs/1508.01991, 2015.

[8] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1064–1074, 2016.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," Proceedings of the 17th North American Chapter of the Association for Computational Linguistics, 2019.

[10] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics, pp.852–860, 2018.

[11] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," Proceedings of the 32nd AAAI Conference on Artificial Intelligence, vol.32, no.1, 2018.

[12] Z. Wu, C. Zheng, Y. Cai, J. Chen, H.-F. Leung, and Q. Li, "Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts," Proceedings of the 28th ACM International Conference on Multimedia, pp.1038–1046, 2020.

[13] K. He, G. Gkioxari, P. Dollár, and R.B. Girshick, "Mask R-CNN," Proceedings of the 16th International Conference on Computer Vision, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Proceedings of the 31st Advances in Neural Information Processing Systems, 2017.

[15] B. Xu, S. Huang, C. Sha, and H. Wang, "MAF: A general matching and alignment framework for multimodal named entity recognition," Proceedings of the The 15th ACM International Conference on Web Search and Data Mining, ed. K.S. Candan, H. Liu, L. Akoglu, X.L. Dong, and J. Tang, pp.1215–1223, 2022.

[16] P.K. Atrey, M.A. Hossain, A. El Saddik, and M.S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," Multimedia systems, vol.16, pp.345–379, 2010.

[17] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," Proceedings of the 23th IEEE International Conference on Information Fusion, pp.1–6, 2020.

[18] C.G.M. Snoek, M. Worring, and A.W. Smeulders, "Early versus late fusion in semantic video analysis," Proceedings of the 13th Annual ACM International Conference on Multimedia, pp.399–402, 2005.

[19] T.C. Dolmans, M. Poel, J.-W.J.R. van't Klooster, and B.P. Veldkamp, "Perceived mental workload classification using intermediate fusion multimodal deep learning," Frontiers in human neuroscience, vol.14, p.609096, 2020.

[20] Y. Wang, Y. Shen, Z. Liu, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," Proceedings of the 33rd AAAI Conference on Artificial Intelligence, vol.33, no.1, pp.7216–7223, 2019.

[21] A.v.d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748, 2018.
[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," International Conference on Machine Learning, 2020.
[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.
[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," Advances in Neural Information Processing Systems, vol.33, pp.18661–18673, 2020.
[25] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics, pp.6894–6910, Nov. 2021.
[26] S.S.S. Das, A. Katiyar, R. Passonneau, and R. Zhang, "CONTaiNER: Few-shot named entity recognition via contrastive learning," Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, Association for Computational Linguistics, pp.6338–6353, May 2022.
[27] Y. Fu, N. Lin, Z. Yang, and S. Jiang, "A dual-contrastive framework for low-resource cross-lingual named entity recognition," arXiv preprint arXiv:2204.00796, 2022.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the 29th Conference on Computer Vision and Pattern Recognition, 2016.
[29] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp.6558–6569, 2019.
[30] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.1990–1999, 2018.
[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol.9, no.8, pp.1735–1780, 1997.
[32] J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.3342–3352, 2020.
[33] D. Chen, Z. Li, B. Gu, and Z. Chen, "Multimodal named entity recognition with image attributes and image knowledge," Proceedings of the 26th Database Systems for Advanced Applications, vol.12682, pp.186–201, 2021.
[34] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity disambiguation for noisy social media posts," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.2000–2008, 2018.

Peng Wang received the B.E. and M.E. degrees from Northwestern Polytechnical University in 2000 and 2003, respectively, and received the Ph.D. degree from Southeast University in 2009. He is an Associate Professor in the School of Computer Science and Engineering, Southeast University. His current research interests include knowledge graph, natural language processing, and social networks.

Xiaohang Chen received the B.E. degree from Soochow University in 2019. He has been a Master student in the School of Computer Science and Engineering, Southeast University, China since 2019. His research interests include natural language processing and multimodal knowledge graph

Ziyu Shang received the B.E. degree from Nanjing University of Science and Technology in 2021. He has been a Master student in the School of Computer Science and Engineering, Southeast University, China. His research interests include knowledge graph, deep learning, and representation learning

Wenjun Ke received the B.E. degree from Northeastern University in 2011, received the M.E. degree from University of Science and Technology of China in 2014, and received the Ph.D. degree from Institute of Computing Technology, Chinese Academy of Sciences, China in 2021. He has been a Senior Engineer in Beijing Institute of Computer Technology and Application, China since 2014. His research interests include natural language processing, opinion mining, and sentiment analysis