Journal of Electronic Research and Application, 2024, Volume 8, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Implicit Modality Mining: An End-to-End Method for Multimodal Information Extraction

Jinle Lu[1], Qinglang Guo[2]*

[1]School of Cyber Science and Technology, University of Science and Technology of China, Hefei 230027, Anhui Province, China
[2]National Engineering Research Center for Public Safety Risk Perception and Control by Big Data (RPP), CETC Academy of Electronics and Information Technology Group Co., Ltd., China Academic of Electronics and Information Technology, Beijing 100041, China

***Corresponding author:** Qinglang Guo, gql1993@mail.ustc.edu.cn

Abstract: Multimodal named entity recognition (MNER) and relation extraction (MRE) are key in social media analysis but face challenges like inefficient visual processing and non-optimal modality interaction. (1) Heavy visual embedding: the process of visual embedding is both time and computationally expensive due to the prerequisite extraction of explicit visual cues from the original image before input into the multimodal model. Consequently, these approaches cannot achieve efficient online reasoning; (2) suboptimal interaction handling: the prevalent method of managing interaction between different modalities typically relies on the alternation of self-attention and cross-attention mechanisms or excessive dependence on the gating mechanism. This explicit modeling method may fail to capture some nuanced relations between image and text, ultimately undermining the model's capability to extract optimal information. To address these challenges, we introduce Implicit Modality Mining (IMM), a novel end-to-end framework for fine-grained image-text correlation without heavy visual embedders. IMM uses an Implicit Semantic Alignment module with a Transformer for cross-modal clues and an Insert-Activation module to effectively utilize these clues. Our approach achieves state-of-the-art performance on three datasets.

Keywords: Multimodal; Named entity recognition; Relation extraction; Patch projection

Online publication: March 30, 2024

1. Introduction

In recent years, social media platforms have seen an increase in user-generated tweets related to events, opinions, preferences, etc [1]. The booming status quo has also created an emerging need to extract structured information from the massive volume of tweets. Named entity identification (NER) [2–4] and relationship extraction (RE) [5–7] are two critical tasks in these rich applications.

Numerous studies have been done on text-based NER and RE tasks [8]. However, in social media scenarios, the form of post content is not only limited to textual modality but also other modalities like images. In this case, information extraction methods that rely solely on text-based models may not be able to extract accurate

information. As shown in **Figure 1**, the text-based named entity recognition model may label bulls as MISC, not ORG, and the text-based relation extraction model cannot determine that Deion Jones and Kwon Alexande Jr are peers. Therefore, multimodal information extraction tasks, especially multimodal named entity recognition (MNER) and multimodal relationship extraction (MNRE), are proposed. Zhang *et al.* [9] proposed a co-attention network to adaptively control and combine text representation with image representation for MNER. Yu *et al.* [10] are the first to apply a unified Transformer structure for the interaction of multimodal information while using a module with auxiliary entity span detection (ESD) to reduce the influence of noisy entities. Ren *et al.* [11] proposed a visual prefix-guided approach to unite text and vision to generate more efficient and robust multimodal representations.

Figure 1. Two examples of multimodal information extraction tasks. (a) MNER with entity label ORG, (b) Multimodal relation extraction with relation peer.

The key to MNER and MNRE tasks is how to effectively incorporate evident visual information to improve textual semantics for NER and NRE tasks. To accomplish this, there are two crucial processes in current state-of-the-art methods. One is the processing of image features, while the other is the interactive mechanism of different modalities. The processing of image features is mainly performed by using a visual grounding toolkit to obtain explicit visual clues like targeted visual objects [12–14]. Then, these visual objects along with the original image are transformed into visual embedding, mainly by using convolutional architectures such as ResNet [15] or a linear embedding layer like Vision Transformers (ViT) [16]. These visual embeddings serve as the output visual representation, facilitating subsequent interactions. This image feature processing is utilized by current state-of-the-art methods [9,10,17]. Regarding modality interaction methods, there is no standard approach. Typically, this involves a combination of self-attention and cross-attention, or sometimes an excessive reliance on the gating mechanism.

Recent studies have demonstrated notable improvements over unimodal models in processing image features and interactive mechanisms, primarily through the adoption of methods such as self-attention and cross-attention. However, despite these advancements, two significant limitations remain unresolved. Firstly, the process of acquiring explicit visual clues via a visual grounding toolkit is burdensome in terms of both time and computational resources. This complexity arises from various components in the pipeline, including the CNN backbone, a region proposal network (RPN), non-maximum suppression (NMS), and region of interest (ROI) head, all of which contribute to increased runtime and computational demands. In academic experiments, the drawbacks associated with using a heavy visual embedder are often overlooked. This is because the acquired visual clues are commonly pre-cached during training to alleviate the computational burden of image feature extraction, rendering them inefficient for real-time reasoning. Additionally, the handling of interactions between different modalities is suboptimal, as it typically relies on alternating self-attention and cross-attention

mechanisms or an excessive dependence on gating mechanisms to leverage explicit visual clues. This approach tends to overly focus on specific visual cues, leading to the misidentification of non-entity clues as entities in images. Consequently, this explicit modeling method may fail to capture nuanced relations between image and text, thereby compromising the model's ability to extract optimal information.

To address the aforementioned challenges, we introduce the Implicit Modality Interaction (IMM) framework. This framework operates solely on the original image, extracting visual embeddings through a linear projection of its patches, thus eliminating the need for a heavy visual grounding toolkit to obtain explicit visual clues. Instead, IMM implicitly captures nuanced relations between image and text, enhancing textual semantics effectively. Specifically, we employ a simple projection of image patches for input, ensuring runtime and parameter efficiency. To facilitate superior interaction in the absence of explicit visual clues, we introduce several modules in our architecture. First is the Implicit Semantic Alignment (ISA) module, which utilizes a Transformer network for modality interactions, incorporating layers of Layer Normalization (LN) [18]. Additionally, we propose a Semantic-Wise (SW) loss, leveraging pseudo-supervised signals from CLIP [19], to align and mine finer-grained information. Furthermore, inspired by prior work [20] suggesting that Knowledge neurons in a feedforward network (FFN) express factual knowledge, we introduce the Insert-Activation (IA) module. IA treats visual clues as knowledge to be activated in FFN, thereby effectively utilizing potentially valid information from the ISA module to enhance textual semantics. In summary, the primary contributions of this paper are as follows:

(1) To the best of our knowledge, our IMM is the first to utilize only linear projection of patches of the original image as visual embedding without any assistance of visual grounding toolkits. This end-to-end design inherently leads to significant runtime and parameter efficiency compared to previous works.

(2) We introduce the ISA module, designed to uncover valuable clues between visual and textual inputs without the need for explicit extraction of visual cues. ISA aims to implicitly mine these cues, enhancing interaction between modalities. Additionally, we propose the IA module, which effectively harnesses the potentially relevant information identified by ISA.

(3) Our experiments, conducted on widely used MNER and MRE datasets, demonstrate that our method achieves new state-of-the-art performance levels. Furthermore, we supplement our findings with ablation studies and case studies, showcasing the pivotal roles played by both the ISA module and the IA module within our framework.

2. Related work

2.1. Multimodal entity and relation extraction

Named entity recognition (NER) and relation extraction (RE) have garnered a lot of interest in the academic community as they are crucial components of information extraction. Early research often involved feature engineering and the utilization of various linear classifiers, including Support Vector Machines (SVM), maximum entropy models, and Conditional Random Fields (CRF) [21,22] However, in recent years, deep learning approaches have shown promising results for NER and RE tasks, employing architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformer models [23–26]. Influenced by the demand for information extraction in realistic social media scenarios, many researchers have focused on multi-modal NER and RE.

In initial works addressing MNER, the primary approach involves encoding text using LSTMs and images using pre-trained CNNs, followed by implicit interaction between the representations of the two modalities.

This methodology has been explored by various researchers. Similarly, in the MRE task, motivated by the challenges observed in MNER, Zheng *et al.* [29] were among the first to propose this task. They demonstrated that traditional text-based RE models perform poorly when applied to social media texts, highlighting the potential benefits of incorporating visual information. Following this, Chen *et al.* [30] utilized graph structure information to align relations between entities in text and images, leveraging image data to supplement missing semantic information. In the current state-of-the-art approaches, both MNER and MRE tasks commonly employ the Transformer Architecture along with visual grounding toolkit methods. Xu *et al.* [31] adopted a strategy where they extract multiple regions from images and utilize a CNN backbone to represent these regions, establishing relationships between image regions and each word in the text. Yu *et al.* [10] introduced a unified Transformer structure for multimodal information interaction, incorporating a module with auxiliary entity span detection (ESD) to mitigate the impact of noisy entities. Devlin *et al.* [32] proposed an alignment and matching framework that employs contrastive learning to enhance the consistency between text and image representations. Chen *et al.* [11] proposed a visual prefix-guided approach to unite text and vision to generate more efficient and robust multimodal representations. Building upon this foundation, we introduce a novel architecture incorporating several essential modules designed to implicitly interact between modalities and extract valuable information for both MNER and MRE tasks.

2.2. Vision-and-language pre-training

Influenced by the success of BERT [33], there has been a growing trend in multimodal research towards Vision-and-Language Pre-training (VLP) on BERT, leading to significant improvements in various downstream multimodal tasks such as visual question answering and image captioning. VLP can be characterized in terms of architecture and pretraining tasks. Architecturally, it can be divided into single-stream structures, including VisualBERT [34], Unicoder-VL [35], VL-BERT [36], and UNITER [37], where patches of image and tokens of text are combined into a sequence and fed into BERT style model to learn contextual embeddings. Alternatively, two-stream structures like LXMERT [38] and ViLBERT [39] process visual and language inputs separately, with interactions facilitated through co-attention or merged attention transformer layers. In terms of pretraining tasks, common approaches include image-text matching (ITM), masked language modeling (MLM), image-text contrastive learning (ITC), and masked region classification (MRC). While these multimodal models have shown consistent improvement in tasks such as image-text retrieval and visual question answering, their application to MNER and MRE may not yield optimal performance. This is because they are typically pretrained on datasets of image captions [40-43], which may not be directly relevant to the objectives of MNER and MRE. We performed several experiments to validate this observation.

3. Methodology

3.1. Overview

In this section, we present a comprehensive overview of our IMM framework designed for multimodal information extraction tasks, as depicted in **Figure 2**. We begin by introducing the ISA module in **Section 3.2.** This module comprised Transformer layers, denoted as LVT layers, shared between the visual and textual sides while maintaining distinct feed-forward layers for each modality. Additionally, we incorporated the CLIP-guided alignment module SW within ISA. Moving forward, in **Section 3.3.**, we detail the IA module. This module consisted of Transformer layers, denoted as LIA layers, featuring separate pathways for visual and textual inputs. Specifically, IA inserts visual representations into the feed-forward layers of the textual pathway, facilitating enhanced integration of visual information with textual semantics.

3.2. Implicit semantic alignment module

3.2.1. Embedding

As shown in **Figure 2**, the input consisted of image-text pairings that provided the appearance characteristics of entity and relations from both the visual and textual modalities. We denoted the text input as t = (*cls*, s_1, …, s_n, sep), where s_1 to s_n represent the token sequence of the input sentence both for sentence-level and word-level. We projected words into token vectors using the pretrained word embedding alignment with transformers architecture [18]:

$$T = [t_{cls}, t_1, \ldots, t_n, t_{sep}] + T_{pos} + T_{pype} \quad (1)$$

where t_{cls} and t_{sep} denoted the start and end tokens, n indicates the length of tokenized subword units, T_{pos} is the position embedding, and T_{type} was the type embedding.

Following the architecture of ViT [16], for the input, where the resolution of the input image is $H \times W$ and C was the number of channels, it is firstly split into $m = H \cdot W/P^2$ patches, where P denoted the patch size, and then linearly projected into patch embeddings :

$$V = [v_{cls}, v_1, \ldots, v_n, v_{sep}] + V_{pos} + V_{pype} \quad (2)$$

where v_{cls} denoted the start token, M indicated the number of patches, V_{pos} was the position embedding, and V_{type} was the type embedding.

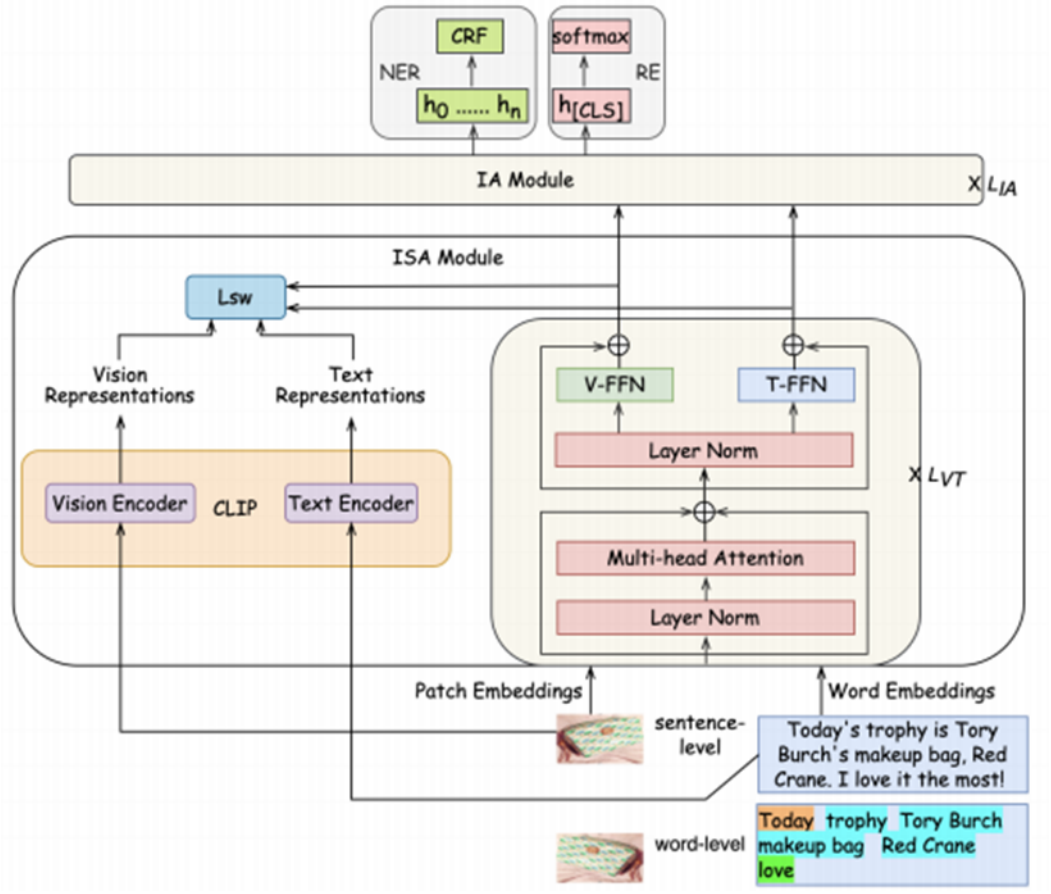

Figure 2. Overall framework of IMM. (1) ISA featuring LVT layers of Transformer shared between the visual and textual sides while retaining respective feedforward layers, along with the CLIP-guided alignment module SW. (2) IA module incorporating LIA layers of Transformer with separate pathways for visual and textual inputs.

3.2.2. Visual-text transformer architecture

Many existing approaches in multimodal information extraction rely on separate models, which have demonstrated effectiveness when utilizing explicit visual clues. However, in our approach, which avoids heavy visual embedding, the absence of cross-modal interactions results in a failure to capture nuanced relations between image and text, thereby undermining the model's ability to extract optimal information. To address this limitation, we introduce the ISA Module, which enhances cross-modal interactions to implicitly capture valuable clues between visual and text inputs.

As depicted in **Figure 2**, the architecture consisted of stacked L_{VT} blocks of Transformer. Within each block, the two modalities share LN and multi-head attention (MHA), facilitating the learning of common spatial mappings between visual and textual modalities. This sharing of parameters aids in understanding the common statistical characteristics of the data. For instance, while LN computes the mean and standard deviation of the input token embeddings, shared LN learns statistically common values across both modalities. From a data-level perspective, this unified cross-modal interaction facilitates the extraction of key information. However, recognizing that visual and textual modalities are inherently different, each block incorporates modality-specific feed-forward layers, denoted as V-FFN and T-FFN modules in **Figure 2**. The processing within each block can be summarized as follows:

$$\mathbf{h}_i^{v/t} = \mathrm{MHA}\left(\mathrm{LN}\left(\mathrm{h}_{i-1}^{v/t}\right)\right) + \mathrm{h}_{i-1}^{v/t} \tag{3}$$

$$\mathbf{h}_i^{v/t} = \mathrm{V/T-FFN}\left(\mathrm{LN}\left(\mathbf{h}_i^{v/t}\right)\right) + \mathbf{h}_i^{v/t} \tag{4}$$

Through L_{VT} layers of Visual-text Transformer Architecture, the model learns a common spatial mapping between the two modalities while preserving their independent characteristics. To enable the model to extract finer-grained information, we facilitated the implicit and automatic mining of valuable information by aligning vision and language at different levels. Specifically, we leveraged CLIP [19], a state-of-the-art multimodal vision and language model comprising a Vision Encoder and a Text Encoder, to obtain vision representations V_c and text representations T_c. Exploiting CLIP's ability to provide signals of image-text similarity, we derived a pseudo-supervised signal P representing the degree of similarity between image-text pairs for alignment:

$$P = \max\left(0, \mathrm{f}(T_{\mathrm{c}}, V_{\mathrm{c}})\right) \tag{5}$$

where f denoted the cosine similarity function. V_c denoted vision representations and T_c denoted text representations at sentence-level. On the interactive side, we simultaneously obtained inputs at both the sentence-level and word-level, with corresponding images being subjected to data augmentation. Inspired by Geva *et al.* [44], we employed the pseudo-supervised signal provided by CLIP to enable the model to gain valuable matching information. By aligning these two different input dimensions, we intended to enable the model to automatically and implicitly mine more precise relations between vision and text at both the sentence-level and word-level. The process is summarized as follows:

$$L_{\mathrm{s}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\left(p_{i,j}^{s}\cdot\log\frac{p_{i,j}^{s}}{q_{i,j}^{s}+\epsilon}\right) \tag{6}$$

$$p_{i,j}^{s} = \frac{\exp(h_i^{ts}\cdot h_j^{vs})}{\sum_{k=1}^{B}\exp(h_i^{ts}\cdot h_k^{vs})} \tag{7}$$

$$q_{i,j}^{s} = \frac{P_{i,j}}{\sum_{k=1}^{B}P_{i,k}} \tag{8}$$

where s denoted sentence-level in all equations, $p_{i,j}^{s}$ denoted the matching probability, h_i^{ts} and h_j^{vs} were the output of L_{VT}th layers, B was the mini-batch size, $q_{i,j}^{s}$ denoted the normalized "true" matching probability. P was defined through Equation 5, with ϵ being a small number to avoid numerical problems. Through the loss L_s we could get some image-text matching information in the output of respective FFN layers. Moreover, to enable the model to automatically and implicitly mine more precise relations between vision and text, we further introduced word-level processing to align the pseudo-supervised signals brought by the sentence-level, intending to extract fine-grained matching information between vision and text. The process is described as follows:

$$L_w = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\left(p_{i,j}^{w}\cdot\log\frac{p_{i,j}^{w}}{q_{i,j}^{w}+\epsilon}\right) \tag{9}$$

$$p_{i,j}^{w} = \frac{\exp(h_i^{tw}\cdot h_j^{vw})}{\sum_{k=1}^{B}\exp(h_i^{tw}\cdot h_k^{vw})} \tag{10}$$

$$q_{i,j}^{w} = q_{i,j}^{s} \tag{11}$$

where w denotes word-level in all equations. $q_{i,j}^{s}$ is defined in Equation 8.

The KL divergence from distribution q top is represented by L_s and L_w. We computed the loss in two directions, i.e., image-to-text and text-to-image, in accordance with prior work [44]. The entire loss is indicated as follows:

$$L_S = \frac{1}{2}(L_s^{t2v} + L_s^{v2t}) \tag{12}$$

$$L_W = \frac{1}{2}(L_w^{t2v} + L_w^{v2t}) \tag{13}$$

$$L_{sw} = \alpha L_S + \beta L_W \tag{14}$$

where α and β are the hyperparameters.

3.3. IA module

By aligning the modal representations and automatically mining implicit information, the ISA module enables us to generate outputs that are rich in information and can be utilized as inputs by the IA module. In this section, we introduce IA, which treats the visual clues as knowledge to be activated in the FFN. This effectively utilizes the potentially valid information brought by the ISA module to enhance textual semantics.

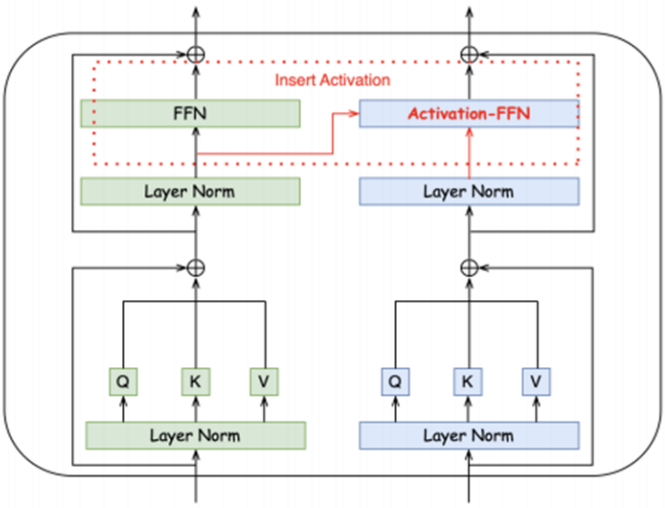

Figure 3. Detailed IA module, which threatens the visual clues as the knowledge to be activated in FFN, that effectively utilizes the potentially valid information brought by ISA module to improve textual semantics.

3.3.1. Insert activation

Previous research [45] has noted that FFN can be thought of as unnormalized key-value memories and emulates brain memory. Specifically, the FFN contained in each layer of the Transformer consists of two linear networks. Through a new perspective, the FFN's computation could be formulated as follows:

$$\mathrm{FFN}\ (\mathbf{h}) = \mathrm{f}\ (\mathbf{h} \cdot K^T) \cdot \mathrm{V} \tag{15}$$

where $\mathbf{h} \in \mathbb{R}^d$ denoted the output of MHA. $K, V \in \mathbb{R}^{d_m \times d}$ were the parameter matrices of the two linear networks, f denoted the activated function like Leaky ReLU [46]; bias terms were ignored. In addition, Wu *et al.* [47] introduced the concept of knowledge-edge neurons, providing initial investigations into how factual knowledge is stored in pre-trained Transformers. More recently, Yao *et al.* [20] successfully utilized knowledge contained in pre-trained language models (PTMs) and external knowledge by transforming it into dense embedding vectors through a knowledge encoder and injecting it into the FFN of the Transformer. Drawing inspiration from this, we consider the hidden embeddings of the visual part as additional knowledge inserted into the text part of the network, to be activated. However, since it has been mentioned earlier that most image regions and text tokens are irrelevant, the information from some images may introduce noise to textual semantics. Therefore, we need to preprocess the inserted neural knowledge to mitigate this issue. Leveraging the output of the ISA module, which contains fine-grained matching information, we conducted token-wise cross-modal interaction as the preprocess of inserted neural knowledge before activation operation. Specifically, we denoted $\mathbf{h}^v \in \mathbb{R}^{m \times d}$ and $\mathrm{h}^t \in \mathbb{R}^{n \times d}$ as the inputs of respective FFN, where m and n denoted sequence length of the visual vectors and textual vectors respectively. We computed the similarity matrix, as shown in Equation 16.

$$S = h^t\ (\ h^v\ )^T \tag{16}$$

Based on Equation 9, the knowledge embeddings were obtained as follows:

$$\mathrm{Sele}_1(h^v) = \mathrm{softmax}(S_i)h^v \tag{17}$$

$$k^v = [\mathrm{Sele}_1\ (x^v); ...; \mathrm{Sele}_n(x^v)] \tag{18}$$

where Agg_i denoted the similarity-aware selected visual representation for i^{th} textual token. denoted the inserted neural knowledge, which would be projected by , for being mapped to the corresponding vector space.

$$\phi_{ik} = \boldsymbol{k}^v \cdot W_{ik} \tag{19}$$

$$\phi_{iv} = \boldsymbol{k}^v \cdot W_v \tag{20}$$

$$Activation - FFN\ (\ h_i^t) = f(\ h_i^t \cdot [\phi_{ik}{:}\boldsymbol{K}_i]) \cdot [\phi_{iv}{::}\boldsymbol{V}_i\ ] \tag{21}$$

In Equation 19, i denotes the i^{th} blocks in IA Module, in which the total number of its blocks are L_{LA} as shown in **Figure 2**.

3.4. Classifier

3.3.1. NER head

To increase model capacity and interaction frequency, we stacked layers of the L_{IA} IA module to form a cascaded architecture. Ultimately, we considered only the word representations of the encoding output, denoted as $H_W \in \mathbb{R}^{n \times d}$, which were then fed into the decoding layer for sequence labeling. Recognizing the dependencies between successive labels, we jointly modeled the hidden representations using a standard CRF layer. Denoting Y' as the set of all possible label sequences for the input sentence X, the probability of the label sequence Y could be calculated using Equation 22.

$$p(Y \mid H_w) = \frac{\prod_{i=1}^{n} \psi_i(y_{i-1}, y_i, H_W)}{\sum_{y' \in Y'} \prod_{i=1}^{N} \psi_i(y'_{i-1}, y'_i, H_W)} \quad (22)$$

where $\psi_n(y_{i-1}, y_i, H_W) = \exp(W_{crf} H_w + b_{crf})$ represents the scoring function, and W_{crf} and b_{crf} are the weight vector and bias. The objective of NER is defined as follows:

$$L_{\text{ner}} = -\sum_{i=1}^{n} \log(p(y^{(i)} \mid f(x^i))) \quad (23)$$

3.4.2. RE head

The goal of RE head was to predict the relation $R \in Y$ between subject entity and object entity. Specifically, a [CLS] head was utilized to compute the probability distribution over the class set Y with the softmax function $p(R \mid X) = \text{softmax}(\mathbf{W}\mathbf{H}^{L}_{[CLS]})$, and the parameters of L and W were fine-tuned by minimizing the cross-entropy loss over $p(R \mid X)$ on the entire X, as shown in Equation 24.

$$L_{re} = -\sum_{i=1}^{n} log(p(R^{(i)} \mid x^i)) \quad (24)$$

IMM was trained by minimizing loss L_{NER} or L_{RE} as follows:

$$L_{NER} = \gamma L_{SW} + \delta L_{ner} \quad (25)$$

$$L_{RE} = \gamma L_{SW} + \delta L_{re} \quad (26)$$

where γ and σ were hyperparameters.

4. Experiments and discussion

This section describes our IMM in MNER and MRE experimental settings. Results obtained on three datasets demonstrate that our IMM framework outperforms other baselines, including both unimodal and multimodal approaches.

4.1. Datasets

For MNER, we conducted experiments on two publicly available Twitter datasets: Twitter-2015 [9] and Twitter-2017 [27]. For MRE, we evaluated our approach on the MNRE dataset [29], which is a manually-labeled dataset specifically curated from Twitter for multimodal relation extraction tasks.

4.2. Metrics

Aligning with other works [9,10,48,49], we used F1 score (F1), precision (P), and recall (R) to evaluate the performance of MNER and MRE.

4.3. Compared baseline

To show the superiority of our IMM, we conducted a comprehensive comparison with several baseline models. Firstly, to illustrate the improvement achieved by incorporating visual information, we compared IMM with traditional text-based models. Secondly, we contrasted IMM with multimodal models, which are pre-trained visual-language models and exhibit either a single-stream or two-stream structure. Additionally, we further considered another group of previous state-of-the-art multimodal approaches for both MNER and MRE tasks.

MNER baselines contain the following approaches: (1) BiLSTM-CRF [2] utilizes word- and character-level

representations via BiLSTM and CNN for NER; (2) CNN-NER [4] is a Twitter-specific NER system with various features to boost performance; (3) AdapCoAtt-BERT-CRF [9] designs an adaptive co-attention network to induce word aware visual representations for each word; (4) UMT [10] extends the Transformer to a multi-modal version and incorporates the auxiliary entity span detection module; (5) UMGF [50], allows a unified multimodal graph fusion approach for MNER and achieves the newest SOTA for MNER.

The MRE baselines involve the following approaches: (1) PCNN [51] uses convolutional networks with piecewise pooling; (2) MTB [52] is a RE-oriented pre-training model based on BERT; (3) Chen *et al.* [30] proposed BERT+SG for MRE, which concatenate the textual representation from BERT with visual features generated with scene graph (SG); (4) MEGA [30] designs the dual graph alignment of the correlation between entities and objects; 5) MKGformer [17] presents a hybrid Transformer network for multimodal tasks, which is the newest state-of-art for MRE.

4.4. Overall performance

Table 1 and **Table 2** show the final model performances upon MNER and MRE. From the experimental results, the following observations were obtained:

(1) Performances are indeed improved by visual information. We found that previous multimodal approaches could achieve better performance, the enormous improvement of F1 score for NER was about 1.7% (comparing UMT with BERT-CRF) and about 5.55% for RE (comparing MEGA with MTB). Hence by comparing the previous state-of-art multimodal techniques with their respective text-based baselines, we can conclude that the visual elements are generally beneficial for MNER and MRE tasks.

(2) Pre-trained multimodal models hold poor performance. To further perform comparative tests, we altered the standard pre-trained vision-language model VisualBERT and ViLBERT with [CLS] classifier for the MRE task and CRF classifier for the MNER task. We noticed that VisualBERT and ViLBERT performed worse than our method and previous state-of-art multimodal approaches. Upon analysis, the pre-trained datasets and objects contained gaps in information extraction tasks, which could be the cause of the pre-trained multimodal poor performance.

(3) Our proposed IMM achieves the best results upon two tasks. Our results (**Table 1**) showed that our IMM outperformed the most recent SOTA model, MKGformer, which enhance F1 scores by 0.69% for the Twitter-2015 dataset and 1.06% for the MNRE dataset. These results indicate that our method can achieve the best performance by utilizing only linear projection of patches of the original image as visual embedding without any assistance of visual grounding toolkit, and this end-to-end design indeed leads to significant runtime and parameter efficiency compared to previous works.

Table 1. Overall performance comparison of the different competitive baseline approaches for MRE

Modality	Methods	MNRE			Twitter-2015			Twitter-2017		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Text	BiLSTM-CRF	-	-	-	66.24	68.09	67.15	80.00	78.76	79.37
	CNN-NER	-	-	-	70.32	68.05	69.17	82.69	78.16	80.37
	BERT-CRF	-	-	-	69.22	74.59	71.81	83.32	83.57	83.44
	PCNN	62.85	49.69	55.49	-	-	-	-	-	-
	MTB	64.46	57.81	60.86	-	-	-	-	-	-

Table 1 (Continued)

Modality	Methods	MNRE			Twitter-2015			Twitter-2017		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Text + Image	AdapCoAtt-BERT-CRF	-	-	-	69.87	74.59	72.15	85.13	83.20	84.10
	UMT	62.93	63.88	63.46	71.67	75.23	73.41	85.28	85.34	85.31
	BERT+SG	62.95	62.65	62.80	73.71	71.21	72.92	84.13	83.88	84.00
	MEGA	64.51	68.44	66.41	70.35	74.58	72.35	84.03	84.75	84.39
	VisualBERT	57.15	59.48	58.30	68.84	71.39	70.09	84.06	85.39	85.04
	ViLBERT	64.50	61.86	63.16	68.23	70.45	69.32	84.62	85.47	85.04
	MKGformer	82.67	81.25	81.95	73.87	76.82	75.32	86.98	**88.01**	87.49
	IMM	**82.58**	**83.45**	**83.01**	**74.51**	**77.69**	**76.01**	87.33	**87.91**	**87.62**

Table 2. Overall performance comparison of the different competitive baseline approaches for MNRE

Modality	Methods	MNRE		
		Precision	Recall	F1
Text	BiLSTM-CRF	-	-	-
	CNN-NER	-	-	-
	BERT-CRF	-	-	-
	PCNN	62.85	49.69	55.49
	MTB	64.46	57.81	60.86
Text + Image	AdapCoAtt-BERT-CRF	-	-	-
	UMT	62.93	63.88	63.46
	BERT+SG	62.95	62.65	62.80
	MEGA	64.51	68.44	66.41
	VisualBERT	57.15	59.48	58.30
	ViLBERT	64.50	61.86	63.16
	MKGformer	**82.67**	81.25	81.95
	IMM	82.58	**83.45**	**83.01**

4.5. Ablation study

We further conducted an ablation study to prove the effect of different modules in our IMM. (1) *w/o SW* refers to the model without the assistance of CLIP for implicit alignment of vision and language at various levels; (2) *w/o IA* refers to the model without the IA module; 3) *w/o SW* and *IA* refers to the model without either of the two modules. **Figure 4** presents the comprehensive experimental data. Notably, the ablation models exhibited a decline in performance, indicating the efficacy of individual elements in our methodology. The following observations were made:

(1) The effectiveness of SW or IA. The SW module and IA module are core components within our IMM framework. The SW module serves as a pluggable operation that harnesses pseudo-supervised signals through CLIP. As evidenced by the experimental results depicted in **Figure 4**, this operation proves to

be effective in providing a priori signaling. On the other hand, the IA module aims to efficiently utilize potentially valid information brought by downstream modules to enhance textual semantics. Similarly, the experimental findings shown in **Figure 4** validate the effectiveness of this approach in leveraging information from downstream modules.

(2) The necessity of combining SW and IA. aims to design an end-to-end structure that inherently achieves significant runtime and parameter efficiency compared to previous works. It is the first to utilize only linear projection of patches of the original image as visual embedding without the assistance of a visual grounding toolkit. To achieve this, we first proposed the ISA module with SW to compensate for the absence of a visual grounding toolkit. Subsequently, to further leverage the information brought by downstream modules, we introduced the IA module. Ultimately, our experimental results demonstrate the effectiveness of the combination of SW and IA in enhancing the performance of our framework.

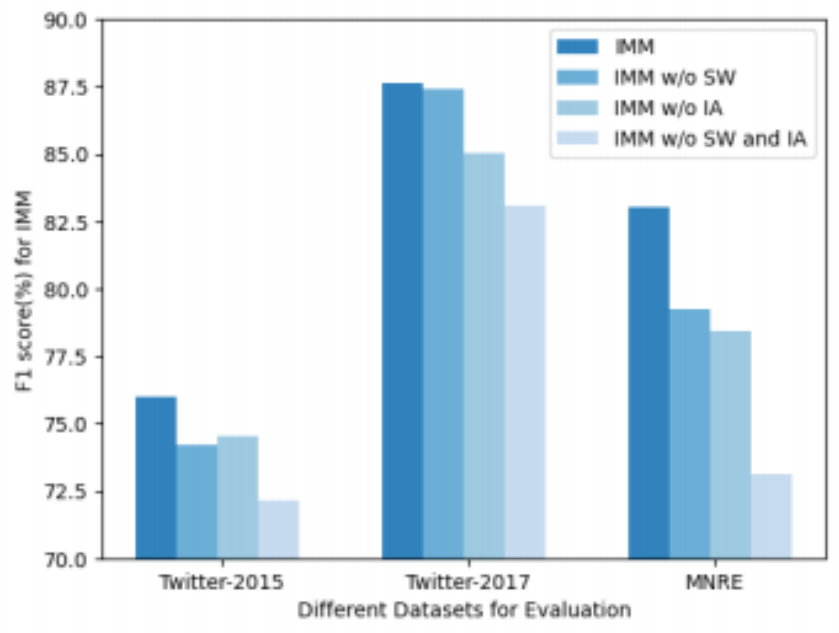

Figure 4. Ablation study results of our IMM

4.6. Cross-task scenario

Table 3 presents a performance comparison between our proposed method, IMM, and previous approaches in a cross-task scenario to analyze its versatility. In the first section, "Twitter-2017 → MNRE" indicates that the model trained on Twitter-2017 is subsequently employed for training and testing on the MNRE dataset. Conversely, in the second section, "MNRE → Twitter-2017" denotes the utilization of the MNRE-trained model for further training and evaluation on Twitter-2017. From the results tabulated in this table, it is evident that our IMM significantly outperforms other methods by a considerable margin. This suggests that our approach can achieve further improvement in cross-task settings, demonstrating the effectiveness of a unified task framework. In contrast, methods such as UMGF exhibit inferior performance compared to their previous results on the respective datasets. Moreover, when compared to the recent HVPNeT method, our IMM achieves even greater enhancement. These findings underscore the high potential of our IMM in cross-task scenarios and validate the efficacy of a unified task framework. Therefore, leveraging more image-text data could facilitate the learning of better modality fusion parameters. Furthermore, extending our work to multi-task learning or multi-modal pre-training represents promising research directions for future exploration.

Table 3. Performance comparison of our IMM and others in cross-task scenario

Methods	Twitter-2017 → MNRE	MNRE → Twitter-2017
UMGF	63.85 → 62.90↓ (0.95)	85.51 → 84.35↓ (1.16)
HYPNeT	81.85 → 82.50↑ (0.75)	86.87 → 87.31↑ (0.26)
IMM	83.01 → 83.80↑ (0.79)	87.62 → 87.94↑ (0.32)

5. Conclusion

In this paper, we introduced a novel end-to-end approach, the IMM, designed for MNER and MRE. Specifically, we proposed an ISA module to implicitly mine valuable clues between visual and text modalities instead of explicitly obtaining visual clues. Additionally, we introduce the IA module to effectively utilize potentially valid information brought by ISA, facilitating superior interaction even with a simple visual embedder and enhancing the incorporation of visual information to improve textual semantics. Extensive experiments conducted on widely used MNER and MRE datasets demonstrate that our method achieves new state-of-the-art performance. In the future, for broader multimedia analysis, we aim to adapt our IMM framework to various multimodal tasks.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Moon S, Neves L, Carvalho V, 2018, Multimodal Named Entity Recognition for Short Social Media Posts. arXiv. https://doi.org/10.48550/arXiv.1802.07862

[2] Huang Z, Xu W, Yu K, 2015, Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv. https://doi.org/10.48550/arXiv.1508.01991

[3] Lample G, Ballesteros M, Subramanian S, et al., 2016, Neural Architectures for Named Entity Recognition. arXiv. https://doi.org/10.48550/arXiv.1603.01360

[4] Gui T, Ma R, Zhang Q, et al., 2019, CNN-Based Chinese NER with Lexicon Rethinking. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 4982–4988.

[5] Liu C, Sun W, Chao W, et al., 2013, Proceedings of the Advanced Data Mining and Applications 9th International Conference, December 14–16, 2013: Convolution Neural Network for Relation Extraction. Hangzhou, 231–242.

[6] Zhang D, Wang D, 2015, Relation Classification Via Recurrent Neural Network. arXiv. https://doi.org/10.48550/arXiv.1508.01006

[7] Zhou P, Shi W, Tian J, et al., 2016, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), August 17–22, 2022: Attention-based Bidirectional Long Short-Term Memory Networks for Relation Classification Using Knowledge Distillation from BERT. Calgary, 207–212.

[8] Nayak T, Majumder N, Goyal P, et al., 2021, Deep Neural Approaches to Relation Triplets Extraction: A Comprehensive Survey. Cognitive Computation, 13: 1215–1232.

[9] Zhang Q, Fu J, Liu X, et al., 2018, Proceedings of The AAAI Conference on Artificial Intelligence, February 2–7, 2018: Adaptive Co-Attention Network for Named Entity Recognition in Tweets. New Orleans, 5674–5681.

[10] Yu J, Jiang J, Yang L, et al., 2020, 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, 3342–3352.

[11] Chen X, Zhang N, Li L, et al., 2022, Good Visual Guidance Makes a Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. arXiv. https://doi.org/10.48550/arXiv.2205.03521

[12] Ren S, He K, Girshick R, et al., 2015, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems 28 (NIPS 2015), 1–9.

[13] He K, Gkioxari G, Dollár P, et al., 2017, Proceedings of the IEEE International Conference on Computer Vision, October 22–29, 2017: Mask R-CNN. Venice, 2961–2969.

[14] Yang Z, Gong B, Wang L, et al., 2019, Proceedings of the IEEE/CVF International Conference on Computer Vision, October 27–November 2, 2019: A Fast and Accurate One-Stage Approach to Visual Grounding. Seoul, 4683–4693.

[15] He K, Zhang X, Ren S, et al., 2016, 2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27–30, 2016: Deep Residual Learning for Image Recognition. Las Vegas, 770–778.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2020, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. https://doi.org/10.48550/arXiv.2010.11929

[17] Chen X, Zhang N, Li L, et al., 2022, Hybrid Transformer with Multi-Level Fusion for Multimodal Knowledge Graph Completion. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 904–915.

[18] Vaswani A, Shazeer N, Parmar N, et al., 2017, 31st Conference on Neural Information Processing Systems (NIPS 2017), December 4–9, 2017: Attention is All You Need. Long Beach, 1–11.

[19] Radford A, Kim JW, Hallacy C, et al., 2021, Proceedings of the International Conference on Machine Learning, July 18–24, 2021: Learning Transferable Visual Models from Natural Language Supervision. Virtual, 8748–8763.

[20] Yao Y, Huang S, Dong L, et al., 2022, Natural Language Processing and Chinese Computing, September 24–25, 2022: Kformer: Knowledge Injection in Transformer Feed-Forward Layers. Guilin, China, 131–143.

[21] Zhou G, Su J, 2005, Machine Learning-Based Named Entity Recognition Via Effective Integration of Various Evidences. Natural Language Engineering, 11: 189–206.

[22] Zhang M, Zhou G, Yang L, et al., 2006, Chinese Word Segmentation and Named Entity Recognition Based on a Context-Dependent Mutual Information Independence Model. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, 154–157.

[23] Luo G, Huang X, Lin CY, et al., 2015, Joint Entity Recognition and Disambiguation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 879–888.

[24] Ma X, Hovy E, 2016, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. arXiv. https://doi.org/10.48550/arXiv.1603.01354

[25] Chiu JP, Nichols E, 2016, Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4: 357–370.

[26] Zhang Z, Wu Y, Zhao H, et al., 2020, Semantics-Aware BERT for Language Understanding. Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), 9628–9635.

[27] Lu D, Neves L, Carvalho V, et al., 2018, Visual Attention Model for Name Tagging in Multimodal Social Media. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1990–1999.

[28] Radford A, Narasimhan K, Salimans T, et al., 2018, Improving Language Understanding by Generative Pre-Training. arXiv. https://doi.org/10.48550/arXiv.2107.86618

[29] Zheng C, Wu Z, Feng J, et al., 2021, Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), July 5–9, 2021: MNRE: A Challenge Multimodal Dataset for Neural Relation Extraction with Visual Evidence in Social Media Posts. Shenzhen, 1–6.

[30] Zheng C, Feng J, Fu Z, et al., 2021, Proceedings of the 29th ACM International Conference on Multimedia, October 20–24, 2021: Multimodal Relation Extraction with Efficient Graph Alignment. Virtual, 5298–5306.

[31] Chen D, Li Z, Gu B, et al., Proceedings of the Database Systems for Advanced Applications: 26th International Conference (DASFAA 2021), April 11–14, 2021: Multimodal Named Entity Recognition with Image Attributes and Image Knowledge. Taipei, 186–201.

[32] Xu B, Huang S, Sha C, et al., 2022, MAF: A General Matching and Alignment Framework for Multimodal Named Entity Recognition. Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 1215–1223.

[33] Devlin J, Chang MW, Lee K, et al., 2018, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805

[34] Li LH, Yatskar M, Yin D, et al., 2019, VisualBERT: A Simple and Performant Baseline for Vision and Language. arXIV. https://doi.org/10.48550/arXiv.1908.03557

[35] Li G, Duan N, Fang Y, et al., 2020, Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. Proceedings of the AAAI Conference on Artificial Intelligence, 11336–11344.

[36] Su W, Zhu X, Cao Y, et al., 2019, VL-BERT: Pre-training of Generic Visual-Linguistic Representations. arXiv. https://doi.org/10.48550/arXiv.1908.08530

[37] Chen YC, Li L, Yu L, et al., Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, August 23–28, 2020: Uniter: Universal Image-Text Representation Learning. Glasgow, 104–120.

[38] Tan H, Bansal M, 2019, Lxmert: Learning Cross-Modality Encoder Representations from Transformers. arXiv. https://doi.org/10.48550/arXiv.1908.07490

[39] Jin D, Pan E, Oufattole N, et al., 2020, What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. arXiv. https://doi.org/2009.13081

[40] Lin TY, Maire M, Belongie S, et al., 2014, Proceedings of European Conference on Computer Vision (ECCV 2014), September 6-12: Microsoft coco: Common Objects in Context. Zurich, 740–755.

[41] Krishna R, Zhu Y, Groth O, et al., 2017, Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision, 123: 32–73.

[42] Ordonez V, Kulkarni G, Berg T, 2011, Proceedings of the 24th International Conference on Neural Information Processing Systems, December 12–15, 2011: Im2Text: Describing Images Using 1 Million Captioned Photographs. Granada, 1143–1151.

[43] Sharma P, Ding N, Goodman S, et al., 2018, Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset for Automatic Image Captioning. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2556–2565.

[44] Zhang Y, Lu H, 2018, Deep Cross-Modal Projection Learning for Image-Text Matching. Proceedings of the European Conference on Computer Vision (ECCV), 686–701.

[45] Geva M, Schuster R, Berant J, et al., 2020, Transformer Feed-Forward Layers Are Key-Value Memories. arXiv. https://doi.org/10.48550/arXiv.2012.14913

[46] Maas AL, Hannun AY, Ng AY, et al., 2013, Rectifier Nonlinearities Improve Neural Network Acoustic Models. Proceedings of the 30th International Conference on Machine Learning, 3.

[47] Dai D, Dong L, Hao Y, et al., 2021, Knowledge Neurons in Pretrained Transformers. arXiv. https://doi.org/10.48550/arXiv.2104.08696

[48] Wu Z, Zheng C, Cai Y, 2020, The 28th ACM International Conference on Multimedia, October 12–16, 2020: Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts. Seattle, 1038–1046.

[49] Wang X, Gui M, Jiang Y, et al., 2022, ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition. arXiv. https://doi.org/10.48550/arXiv.2112.06482

[50] Zhang D, Wei S, Li S, et al., 2021, Multi-Modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. Proceedings of the AAAI Conference 0n Artificial Intelligence, 14347–14355.

[51] Zeng D, Liu K, Chen Y, et al., 2015, Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1753–1762.

[52] Soares LB, FitzGerald N, Ling J, et al., 2019, Matching the Blanks: Distributional Similarity for Relation Learning. arXiv. https://doi.org/10.48550/arXiv.1906.03158