# Learning from Different text-image Pairs: A Relation-enhanced Graph Convolutional Network for Multimodal NER

Fei Zhao
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
zhaof@smail.nju.edu.cn

Chunhui Li
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
lich@smail.nju.edu.cn

Zhen Wu[*]
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
wuz@nju.edu.cn

Shangyu Xing
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
starreeze@foxmail.com

Xinyu Dai
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
daixinyu@nju.edu.cn

## ABSTRACT

Multimodal Named Entity Recognition (MNER) aims to locate and classify named entities mentioned in a (text, image) pair. However, dominant work independently models the internal matching relations in a pair of image and text, ignoring the external matching relations between different (text, image) pairs inside the dataset, though such relations are crucial for alleviating image noise in MNER task. In this paper, we primarily explore two kinds of external matching relations between different (text, image) pairs, i.e., inter-modal relations and intra-modal relations. On the basis, we propose a Relation-enhanced Graph Convolutional Network (R-GCN) for the MNER task. Specifically, we first construct an inter-modal relation graph and an intra-modal relation graph to gather the image information most relevant to the current text and image from the dataset, respectively. And then, multimodal interaction and fusion are leveraged to predict the NER label sequences. Extensive experimental results show that our model consistently outperforms state-of-the-art works on two public datasets. Our code and datasets are available at https://github.com/1429904852/R-GCN.

## CCS CONCEPTS

• **Computing methodologies → Information extraction**; • **Information systems → Multimedia and multimodal retrieval**.

## KEYWORDS

multimodal named entity recognition, graph convolutional network, multi-head attention, conditional random field

---

Figure 1: An example of multimodal tweets. In this tweet, "Alibaba" is the name of a Person instead of an Organization.

## 1 INTRODUCTION

Named Entity Recognition (NER) is a subtask of information extraction, which aims to identify text spans to specific entity types such as Person (PER), Location (LOC) and Organization (ORG). NER has been widely used in many downstream tasks such as entity linking [7] and relation extraction [34].

Recently, most of the studies on NER solely rely on text modalities to infer labels [4, 17, 19]. However, when the texts contain polysemy entities, it is really difficult to recognize named entities accurately only depending on textual information [18, 20]. One promising solution is to introduce other modalities (e.g., image) as the supplement of the textual modality. Take Figure 1 as an example, the word "Alibaba" appearing in tweets could be identified as multiple types of entities such as "Organization" and "Person", but

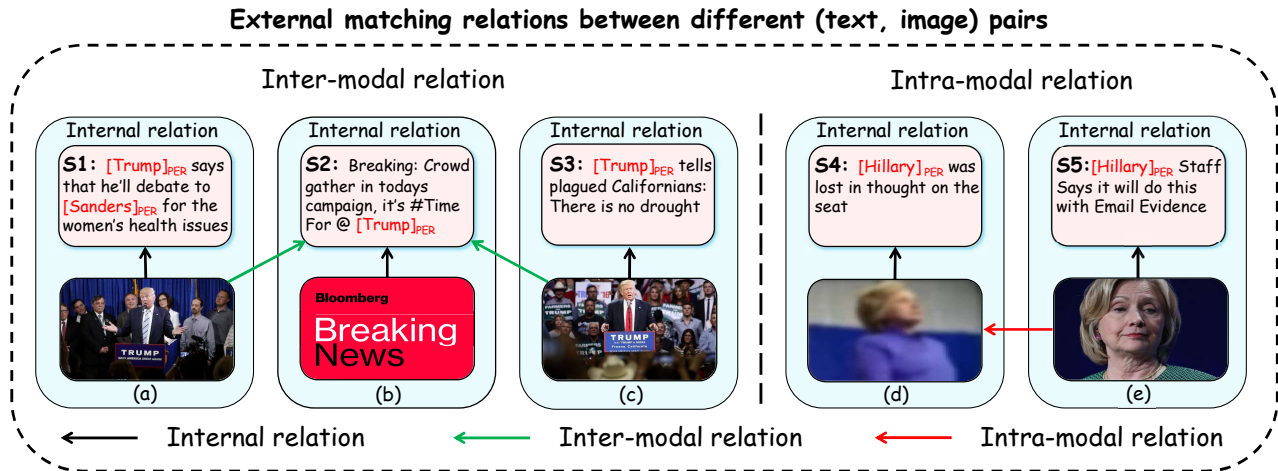## External matching relations between different (text, image) pairs



**Figure 2: Each blue box contains a pair of image and text in the dataset. Named entity and their corresponding entity type are highlighted in the text. The black arrow represents the internal matching relation in a image-text pair. The green arrow represents the inter-modal relation between the text and image in different image-text pairs, and the red arrow represents the intra-modal relation between images in different image-text pairs. Previous studies primarily focus on the internal matching relations between the pair of text and image in a single sample, paying no attention to two kinds of external matching relations between different (text, image) pairs, which is essential for alleviating image noise in MNER task.**

when the word "Alibaba" is aligned with the visual object *person* in the image, "Organization" will be filtered out.

From the above example, we can conclude that aligning the words in the text with the visual objects in the image lies at the heart of multimodal NER (MNER) task. To this end, a lot of efforts have been made, roughly divided into three aspects: (1) encoding the whole image into a global feature vector, and then design effective attention mechanisms to extract the visual information related to the text [18]; (2) segmenting the whole image averagely into multiple visual regions, and then explicitly model the relevance between the text sequence and visual regions [2, 20, 24, 25, 33, 38]; (3) only retaining the visual object regions in the image, and then make them interact with the text sequence [30, 31, 36, 40].

Despite their promising results, the above studies independently model the internal matching relations in a pair of image and text, ignoring the external matching relations between different (text, image) pairs. In this work, we argue that such external relations are crucial for alleviating image noise in MNER task. Especially, we explore two kinds of external matching relations inside a dataset:

- **Inter-modal relation:** From the perspective of text, a piece of text may be associated with multiple images inside the dataset. When the named entity in the text does not appear in the corresponding image, other relevant images are generally helpful to identify the named entity in the text. As shown in Figure 2(b), the named entity "Trump" in sentence S2 does not appear in the corresponding image, so it is relatively difficult to infer the named entity tag only depending on informal sentence S2. However, when taking other images closely related to sentence S2 into consideration (e.g., the images in Figure 2(a) and 2(c)), the named entity tag in sentence S2 is largely possible to be "PER" since these relevant images all contain the visual object *person*. Therefore, a

feasible and natural approach should make the connection between text and image in different (text, image) pairs.

- **Intra-modal relation:** From the perspective of images, different images often contain the same type of visual object, the clear visual object region is easier to recognize the named entity than the fuzzy visual object region. For instance, both the images in Figure 2(d) and 2(e) involve a visual object *person*. Although it is relatively difficult to infer the named entity tag in sentence S4 through the fuzzy visual object region in Figure 2(d), we can infer that the named entity tag in sentence S4 is more likely to be "PER" according to Figure 2(e). This is because it is easier to infer named entity tag "PER" through the clear visual object region in Figure 2(e). Therefore, a feasible and natural approach should make the connection between images in different (text, image) pairs.

To well model the above two kinds of external matching relations, we propose a Relation-enhanced Graph Convolutional Network (*R-GCN*) for the MNER task. Specifically, *R-GCN* mainly consists of two modules: The first module constructs an inter-modal relation graph and an intra-modal relation graph to severally gather the image information most relevant to the current text and image from the dataset. The second module performs multimodal interaction and fusion followed by predicting the NER label sequences. Extensive experimental results show that our network consistently outperforms state-of-the-art works on two public datasets.

Our contributions are summarized as follows: (1) As far as we know, we are the first to propose leveraging the external matching relations between different (text, image) pairs to improve the performance on the MNER task. (2) We design a Relation-enhanced Graph Convolutional Network (*R-GCN*) to model inter-modal relations and intra-modal relations simultaneously. (3) Experimental

results on two public datasets achieve state-of-the-art performance. Further analysis verifies the validity of our proposed network.

## 2 RELATED WORK

Named Entity Recognition has been received more and more attention in the past few years due to its wide application in the industry [1, 23, 27, 35]. A traditional way is to perform NER on pure text with RNN feature extractors and sequence labeling techniques. Recently, visual modality is suggested to be added to texture modality for a better result [3], since in many scenarios images appear alongside with text (e.g., social media and shopping reviews). However, how to leverage visual information and incorporate those two modalities remains a non-trivial task, and researches in this field are relatively limited. In our work, Graph Neural Networks (GNN) are employed to model external matching relations between different (image, text) pairs. Therefore, in the following, we give a brief overview of NER, MNER and GNN respectively.

### 2.1 Named Entity Recognition

Traditionally, early studies handle NER tasks with feature engineering and different linear classifiers such as SVM or maximum entropy, and they rely on CRF for sequence labeling [37, 41]. Some even require a external auxiliary knowledge database and handcrafted features [13]. Later, neural networks are employed in NER tasks such as LSTM, CNN and attention mechanism [12, 19, 32]. Recently, structures based on large-scale pre-trained language models such as BERT have produced even better outcomes [39].

However, none of the above takes visual modality into consideration, which is a supplement supposed to be helpful in improving performance furthermore, especially in scenarios where text is short or the dataset is small [3].

### 2.2 Multimodal Named Entity Recognition

Typically, studies on MNER are similar in textual feature extraction, but each employs different techniques when leveraging image information and fusing the two modalities. Overall, their work can be divided into three categories:

1) **Encoding the whole image into a global feature vector, and then design effective attention mechanisms to extract visual information related to the text.** For example, Moon et al. [20] view word embedding, character embedding and global image vector (encoded by a simple CNN network) as three independent modalities, and used Modality Attention to get fused representations.

2) **Segmenting the whole image averagely into multiple visual regions, and then model the relevance between the text sequence and visual regions.** Lu et al. [18] used the pre-trained model ResNet [10] to extract visual regions, and then added them to text embedding through visual attention model. Zhang et al. [38] employed adaptive co-attention Network to incorporate textual and visual regions. Yu et al. [33] proposed multimodal interaction module to integrate the two modalities and entity span detection module to help filter out visual noise. Chen et al. [2] leveraged the external knowledge database and attention-guided visual layer to obtain the final multimodal representation. Sun et al. [25] and Sun

et al. [24] used modified BERT encoder to obtain the fused representation, and then introduced text-image relation classification as a subtask to decide whether the image feature is useful.

3) **Only retaining the visual object regions in the image, and then make them interact with the text sequence.** Wu et al. [30] employed Faster-RCNN [21] to extract object region features which is fed into adaptive co-attention network. Wu et al. [31] resort to Mask-RCNN [9] for object detection, and they embedded top-k objects into vectors to interact with text features through dense co-attention Layer. Zheng et al. [40] used adversarial learning technique, aiming at fusing textual and visual features into a common feature space. Zhang et al. [36] first extracted noun phrases in text and visual object regions in images, and then employed GNN to model the relations between them.

However, all of the studies above focus on the internal matching relations between the pair of text and image in a single sample, paying no attention to external matching relations in different pairs, which is crucial for alleviating image noise in MNER task.

### 2.3 Graph Neural Networks

GNN is designed to model complex graph structures [8], whose variants such as graph convolutional network [16] and graph attention network [29] have shown promising results on a wide range of tasks. Zhang et al. [36] first introduced GNN into MNER task, in which they regard noun phrases in text and visual object regions in images as nodes and used different edges to mark the object-to-object and noun-to-noun relations in a (text, image) pair.

Different from their work in which graph is constructed based on a single sample, we create our GNN on the entire dataset. This enables us to model relations between different (text, image) pairs.

## 3 METHODOLOGY

In this section, we first formulate the MNER task, and then give an overview of our model.

**Task Definition** Given a sentence $S$ and its associated image $I$, the goal of MNER is to locate and classify named entities mentioned in the sentence into predefined semantic categories, e.g., Person (PER), Location (LOC), Organization (ORG), and Miscellaneous (MISC). As with previous work, we formulate this task as a sequence labeling problem. Specifically, let $S = (w_1, w_2, ..., w_n)$ denote a sequence of input words, and $y = (y_1, y_2, ..., y_n)$ be the corresponding label sequence, where $y_i \in Y$ and $Y$ is the pre-defined label set with the BIO2 tagging schema [22].

**Overview** In this paper, we propose a Relation-enhanced Graph Convolutional Network (*R-GCN*) to model two kinds of external matching relations respectively. Figure 3 shows the overall architecture of the *R-GCN* model which consists of five major modules: 1) Feature Extraction Module; 2) Inter-Modal Relation Module; 3) Intra-Modal Relation Module; 4) Multi-modal Interaction Module; 5) CRF Decoding Module. In the following, we will illustrate the five main components of *R-GCN* model respectively.

### 3.1 Feature Extraction Module

*3.1.1 Text Representation.* As a strong text encoder, pre-trained model BERT [5] is widely used in different NLP tasks. Here, we
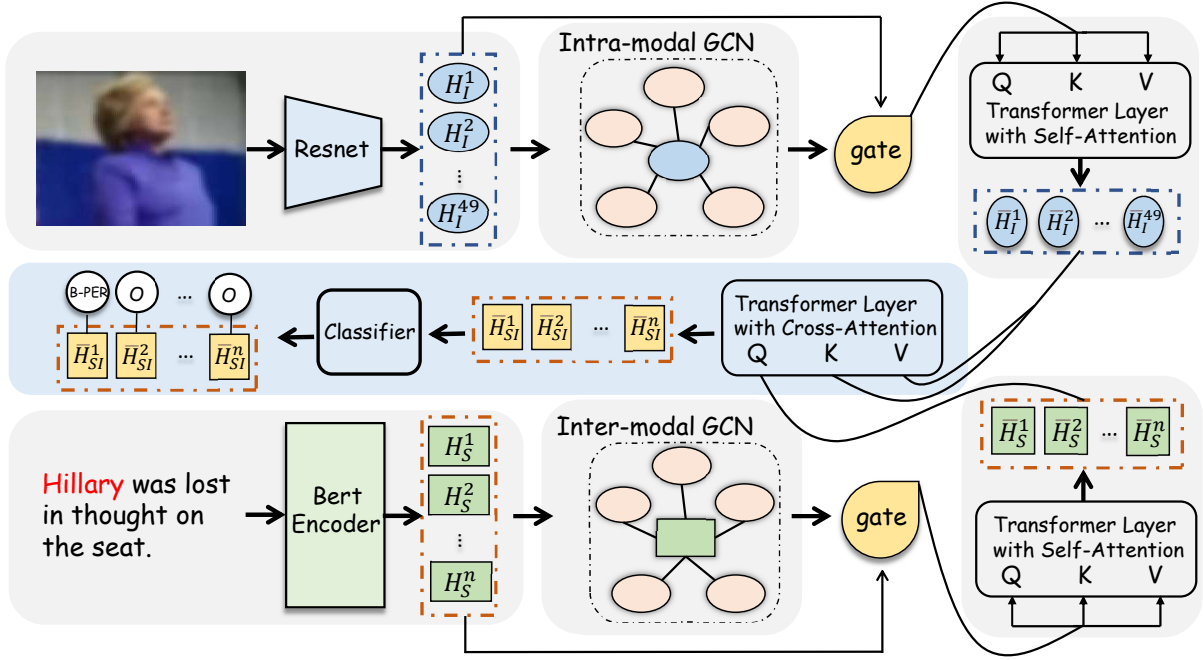
**Figure 3: The overview of our proposed *R-GCN* model.**

input the sentence $S$ into the BERT-based model to obtain the contextualized representations as follows:

$$H_S = BERT(S), \qquad (1)$$

where $H_S \in \mathbb{R}^{n \times d}$ denotes the text representations, $d$ is the hidden dimension, and $n$ is the length of the sentence.

*3.1.2 Image Representation.* Most of the previous methods in MNER task use the pre-trained model ResNet [10] to extract image features. For a fair comparison, our image encoder should be the same as previous models. Specifically, given an image $I$, we first resize $I$ to 224×224 pixels, and then adopt image recognition model ResNet-152 [10] to obtain the output of the last convolutional layer:

$$\text{ResNet}(I) = \{r_j | r_j \in \mathbb{R}^{2048}, j = 1, 2, ..., 49\}, \qquad (2)$$

which splits the original image into $7 \times 7 = 49$ regions and each region is represented by a 2048-dimensional vector $r_j$. Next, we project the visual features to the same space of textual features:

$$H_I = W_v \text{ResNet}(I), \qquad (3)$$

where $W_v \in \mathbb{R}^{d \times 2048}$ is the learnable parameter.

## 3.2 Inter-Modal Relation Module

According to our observations, a piece of text may be associated with multiple images inside the dataset. When the named entity in the text does not appear in the corresponding image, other relevant images are generally helpful to identify the named entity in the text. To this end, we propose an inter-modal relation graph to integrate other images in the dataset that have a similar meaning to the input sentence. In the following, we first present how to

build the nodes and edges of the inter-modal relation graph, and then describe our method in detail.

- **Nodes**: There are two types of nodes in the inter-modal relation graph, namely text node and image node. The text node serves as the central node, which is obtained by feeding the current sentence $S$ to the BERT encoder, as is stated in Eq.1, while the image node is the image representation extracted from pre-trained model ResNet, aiming to provide auxiliary information to the central text node.
- **Edges**: Our goal is to measure whether other images in the dataset contain similar scenes mentioned in the sentence $S$. However, it's not trivial to achieve due to the semantic gap between the image and text. To this end, we first exploit the image caption model [14] to transform the image into the caption description, and then regard the cos similarity between the input sentence $S$ and caption description as an edge of text node and image node.

Specifically, we first employ the pre-trained image caption model (CATR) [14] to obtain caption description for each image $P$ in the dataset, and then input it into the BERT encoder to obtain the contextualized representation:

$$T_P = BERT(CATR(P)), \qquad (4)$$

where $T_P \in \mathbb{R}^{d \times n}$ denotes the caption representation, $d$ is the hidden dimension and $n$ is the length of caption description.

After that, except for paired image $I$ with the sentence $S$, we calculate the cos similarity between the text representation $H_S$ in Eq.1 and the caption representation $T_P$ of other images in the dataset, so as to find out Top-$K$ images that express a similar meaning with

the input sentence $S$. Formally,

$$M \in \Omega(S) := [cos(H_S, T_P)]_K, \tag{5}$$

where $M \in \Omega(S)$ denotes a set of similar images of sentence $S$, $[\cdot]_K$ is used to pick $K$ images with the highest similarity score. After obtaining the Top-$K$ images, we construct an inter-model relation graph by taking the image representation or text representation as a node and the cos similarity between the text representation of input sentence and the caption representation of images as an edge. Finally, we do convolutions on every text node with its neighboring Top-$K$ image nodes in the graph. Formally,

$$H_S^{GCN} = \sigma(\sum_{H_M \in \Omega(H_S)} W_S H_M + b_S), \tag{6}$$

where $H_M \in \Omega(H_S)$ denotes the set of similar image representation of text representation $H_S$, $W_S$ is a linear transformation weight, $b_S$ is a bias term, and $\sigma$ is a nonlinear function.

**Text Gate**: It is inevitable to find wrong similar images from the dataset. To weaken the negative impact of noisy images, we design a text gate to control how much similar images representation will contribute to the text representation. Formally,

$$\hat{H}_S = H_S + \lambda_S H_S^{GCN}, \tag{7}$$

$$\lambda_S = sigmoid(W_S^{GCN}[H_S; H_S^{GCN}]), \tag{8}$$

where $W_S^{GCN}$ is a trainable weight matrix, $[;]$ denotes the concatenation operation, and sigmoid$(\cdot)$ is a nonlinear function.

### 3.3 Intra-Modal Relation Module

As mentioned before, when different images contain the same type of visual objects, the clear visual object region is easier to recognize the named entity in the text than the fuzzy visual object region. To this end, we construct an intra-modal relation graph, so as to gather the similar images which contain the same type of visual objects as input image $I$. In the following, we first present how to build the nodes and edges of intra-modal relation graph, and then describe our method in detail.

- **Nodes**: For each image in the dataset, we regard the image representation extracted from pre-trained model ResNet as an image node, in which the center node is current input image $I$, as is stated in Eq.3.
- **Edges**: Our goal is to measure whether other images in the dataset contain the same types of visual object with input image $I$. Apparently, ResNet has no ability to get visual object regions. Thus, we first exploit the object detection model to obtain a group of visual objects for each image, and then regard the cos similarity between the object representation of input image and other image as an edge of image node.

Inspired by [6], we first use the pre-trained object detection model Faster-RCNN [21] to generate a group of visual objects $V_I^{obj} = \{v_i^{obj}\}_{i=1}^O$ for input image $I$. Then, we average all these object region to obtain the object representation for input image $I$:

$$V_I = \frac{1}{O} \sum_{i=1}^O v_i^{obj} \tag{9}$$

where $O$ denotes the number of object regions in an image and $v_i^{obj}$ means a 2048-dimensional feature of the $i$-th object region. After that, we calculate the cos similarity between the object features $V_I$ of the input image $I$ and the object features $V_N$ of other images in the dataset, so as to find out similar images which contain the same type of visual object as input image $I$. Formally,

$$N \in \Omega(I) := [cos(V_I, V_N)]_K \tag{10}$$

where $N \in \Omega(I)$ indicates a set of similar images with input image $I$, $[\cdot]_K$ is used to pick $K$ images with the highest similarity score. After obtaining the Top-$K$ images, we build an intra-model relation graph by taking the image representation as a node and the cos similarity between the object representation of input image and other image as an edge. Finally, the center node gather the image features of Top-$K$ neighbor nodes in the graph. Formally,

$$H_I^{GCN} = \sigma(\sum_{H_N \in \Omega(H_I)} W_I H_N + b_I) \tag{11}$$

where $H_I$ is the image representation of input image $I$ in Eq. 3, $H_N \in \Omega(H_I)$ denotes a set of similar image representation of $H_I$, $W_I$ is a linear transformation weight, $b_I$ is a bias term.

**Image Gate**: Similarly, we also design an image gate to control how much similar image representation will contribute to the input image representation. Formally,

$$\hat{H}_I = H_I + \lambda_I H_I^{GCN}, \tag{12}$$

$$\lambda_I = sigmoid(W_I^{GCN}[H_I; H_I^{GCN}]), \tag{13}$$

where $W_I^{GCN}$ is a trainable weight matrix.

### 3.4 Multi-modal Interaction Module

Except for the above two kinds of external matching relations between different (image, text) pairs, as with previous work, we also need to model the internal matching relations in a pair of image and text. Following the previous methods [33, 36], we first employ self-attention Transformer layer to enhance the interaction of different words or different visual regions in the single modality (i.e., text or image). And then, we adopt the cross-attention transformer layer to make an interaction between text and image. Formally,

$$\overline{H}_S = \text{Self-ATT}(\hat{H}_S, \hat{H}_S, \hat{H}_S), \tag{14}$$

$$\overline{H}_I = \text{Self-ATT}(\hat{H}_I, \hat{H}_I, \hat{H}_I), \tag{15}$$

$$\overline{H}_{SI} = \text{Cross-ATT}(\overline{H}_S, \overline{H}_I, \overline{H}_I), \tag{16}$$

where Self-ATT$(\cdot)$ denotes self-modal multi-head attention as [28], Cross-ATT$(\cdot)$ denotes the cross-modal multi-head attention as [26], $\overline{H}_S$, $\overline{H}_I$, and $\overline{H}_{SI}$ denotes the final text representation, image representation and word-aware visual representation.

### 3.5 CRF Decoding Module

We apply the Conditional Random Fields (CRF) [17] decoder to perform conditional sequence labeling. CRF considers the correlations between lables in neighborhoods and score the whole sequence of lables. Specifically, we use a linear-chain CRF and score the tag sequence as conditional probability:

$$p(y|\overline{H}_{SI}) = \frac{\prod_{i=1}^N F_i(y_{i-1}, y_i, \overline{H}_{SI})}{\sum_{y' \in Y} \prod_{i=1}^N F_i(y'_{i-1}, y'_i, \overline{H}_{SI})} \tag{17}$$

**Table 1: The statistics of two multimodal Twitter datasets.**

| Entity Type | TWITTER-15 | | | TWITTER-17 | | |
|---|---|---|---|---|---|---|
| | **Train** | **Dev** | **Test** | **Train** | **Dev** | **Test** |
| Person | 2217 | 552 | 1816 | 2943 | 626 | 621 |
| Location | 2091 | 522 | 1697 | 731 | 173 | 178 |
| Organization | 928 | 247 | 839 | 1674 | 375 | 395 |
| Miscellaneous | 940 | 225 | 726 | 701 | 150 | 157 |
| Total | 6176 | 1546 | 5078 | 6049 | 1324 | 1351 |
| Num of Tweets | 4000 | 1000 | 3257 | 3373 | 723 | 723 |

where $F_i(y_{i-1}, y_i, \overline{H}_{SI})$ and $F_i(y'_{i-1}, y'_i, \overline{H}_{SI})$ are potential functions. Finally, we use the maximum conditional likelihood estimation as the loss function of the model, i.e., $L(p(y|\overline{H}_{SI})) = \sum_i log p(y|\overline{H}_{SI})$.

## 4 EXPERIMENTS

### 4.1 Datasets and Evaluation Metrics

To evaluate the effect of Relation-enhanced Graph Convolutional Network (*R-GCN*), we conduct experiments on two public multimodal datasets: TWITTER-2015 and TWITTER-2017 from [18, 38], which include user posts on Twitter during 2014-2015 and 2016-2017, respectively. Table 1 shows the number of entities for each type and the counts of multimodal tweets in the training, development, and test sets of the two datasets.

Following [33], we use Precision (*Pre*), Recall (*Rec*) and F1 Score (*F1*) as our evaluation metrics. Besides, the paired $t$-test is conducted to test the significance between difference approaches.

### 4.2 Implement details

Our *R-GCN* model is implemented by Tensorflow framework with an NVIDIA Tesla V100 GPU. The word representations are initialized with the pre-trained uncased BERT-based model [5], and the visual representations are initialized by a pre-trained ResNet-152 model [10]. We set the maximum length of the sentence input as 98, mini-batch size as 16, hidden dimension as 768, and the number of attention heads as 8.

During training, we train each model for a fixed 25 epochs, and then select the model with the best F1 score on the development set. Finally, we evaluate its performance on the test set. In this period, we use the Adam optimizer [15] to minimize the loss function, where we set the learning rate as 5e-5 and the dropout rate as 0.9 [11]. Besides, we set the hyper-parameter $K$ to be 5 on both datasets. Finally, we report the average performance and standard deviation over 5 runs with random initialization.

### 4.3 Compared Methods

In order to evaluate the effectiveness of our model, we compare it with two groups of baseline systems. The first group includes several classic text-based NER approaches: **BiLSTM-CRF** [12], **CNN-BiLSTM-CRF** [19], **HBiLSTM-CRF** [17], **BERT** [5], **BERT-CRF**. The second group includes several latest multi-modal approaches for MNER task: **ACOA** [38], **VG** [18], **OCSGA** [31], **IAIK** [2], **Object-AGBAN** [40], **UMT** [33], **RpBERT** [25], **UMGF** [36].

## 5 RESULTS AND DISCUSSION

### 5.1 Main Results

We conduct experiments on TWITTER-2015 and TWITTER-2017 datasets. As shown in Table 2, we report the overall Precision, Recall and F1 score, as well as F1 score for each single type. As with previous work, we mainly focus on overall F1 score. Based on these results, we can make a couple of observations:

(1) For text-based methods, pre-trained model *BERT* performs better than the conventional neural networks apparently. We attribute this to the fact that pre-trained model can provide abundant syntactic and semantic features. Besides, in terms of overall results of both datasets, we found that *BERT-CRF* with CRF decoding performs a little better than *BERT* except for the metric *Rec*. A possible reason is that CRF layer consider the dependencies between words.

(2) Multi-modal approaches *ACOA* and *VG* are variants of uni-modal approaches *CNN-BiLSTM-CRF* and *HBiLSTM-CRF*, respectively. It's clear that the former generally perform better than the latter, which indicates that the image information can be used as the supplement of the textual information, and thus improves the performance for MNER task. Besides, *UMT* and *UMGF* perform better than other multi-modal approaches. A possible reason is that they employ self-modal and cross-modal multi-head attention to learn more robust representation than other approaches.

(3) In comparison with *UMT* and *UMGF*, *R-GCN* achieves competitive results on both datasets. It is worth mentioning that our *R-GCN* model outperforms the current state-of-the-art model *UMGF* by 1.48% and 1.97% on overall *F1*-score, respectively. Besides, with regard to single type, *R-GCN* at most outperforms *UMGF* by 1.86% and 5.08% on TWITTER-2015 and TWITTER-2017 datasets. These results further reveal the effectiveness of our model.

(4) *R-GCN w/o Gate* is a variant of *R-GCN*, which remove the combinations of text gate and visual gate. Apparently, *R-GCN w/o Gate* performs slightly worse than *R-GCN* on most setting. The reason behind this may be that *R-GCN* filters out some noisy image information through the gating mechanism.

### 5.2 Ablation Study

To investigate the impacts of individual modules and combinations of several components on the overall effect of the model, we conducts an ablation study on two modules in *R-GCN*. The results are shown in Table 3, where "w/o" indicates the removal of the single components or several components, "InterRG" denotes the Inter-model relation module, and "IntraRG" means Intra-model relation module. Concretely, we can make the following conclusions:

(1) As expected, removing any of the two modules makes the overall performance worse, which validates the rationality of leveraging the external matching relations between different (text, image) pairs inside a dataset to improve the performance of the MNER task. After removing two components at the same time, the performance of the model is further degraded, which demonstrates that *IntraRG* and *InterRG* modules improve the performance of the MNER task from different perspectives.

(2) In comparison to Intra-model relation module (*IntraRG*), Inter-model relation module (*InterRG*) has a greater impact on the performance of *R-GCN*. This is because we primary rely on the text sequence to predict the NER label sequence. Hence, integrating the

**Table 2: Performance comparison on the TWITTER-15 and TWITTER-17 datasets (%). For the baseline model, the results with ♣ are obtained by running the code released by the author, and the other results without symbols are retrieved from the original papers. We report the average performance and standard deviation over 5 runs with random initialization. Best results are in bold. The marker † refers to significant test p-value < 0.05 when comparing with UMT and UMGF.**

| Methods | TWITTER-2015 | | | | | | | TWITTER-2017 | | | | | | |
| | Single Type ($F_1$) | | | | Overall | | | Single Type ($F_1$) | | | | Overall | | |
| | PER | LOC | ORG | MISC | Pre | Rec | F1 | PER | LOC | ORG | MISC | Pre | Rec | F1 |
| Text | | | | | | | | | | | | | | |
| BiLSTM-CRF | 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 | 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 |
| HBiLSTM-CRF | 82.34 | 76.83 | 51.59 | 32.52 | 70.32 | 68.05 | 69.17 | 87.91 | 78.57 | 76.67 | 59.32 | 82.69 | 78.16 | 80.37 |
| BERT | 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 |
| Text+Image | | | | | | | | | | | | | | |
| ACOA | 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 |
| VG | 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 |
| OCSGA | 84.68 | 79.95 | 56.64 | 39.47 | 74.71 | 71.21 | 72.92 | – | – | – | – | – | – | – |
| Object-AGBAN | 84.75 | 79.41 | 58.31 | 40.72 | 74.13 | 72.39 | 73.25 | – | – | – | – | – | – | – |
| IAIK | 84.28 | 79.42 | 58.97 | 41.47 | **74.78** | 71.82 | 73.27 | – | – | – | – | – | – | – |
| RpBERT♣ | 85.18 | 81.19 | 58.68 | 37.88 | 71.15 | 74.30 | 72.69 | 89.05 | 84.03 | 82.60 | 63.67 | 82.85 | 84.38 | 83.61 |
| UMT | 85.24 | 81.58 | 63.03 | 39.45 | 71.67 | 75.23 | 73.41 | 91.56 | 84.73 | 82.24 | 70.10 | 85.28 | 85.34 | 85.31 |
| UMT♣ | 84.74 | 81.69 | 60.59 | 39.22 | 72.66 | 74.14 | 73.39 | 90.41 | 83.98 | 81.20 | 65.56 | 84.02 | 84.09 | 84.05 |
| UMGF | 84.26 | **83.17** | **62.45** | **42.42** | 74.49 | 75.21 | 74.85 | 91.92 | 85.22 | 83.13 | 69.83 | 86.54 | 84.50 | 85.51 |
| UMGF♣ | 84.50 | 81.54 | 60.72 | 40.57 | 72.47 | 74.60 | 73.52 | 91.14 | 84.24 | 83.23 | 67.30 | 85.30 | 84.99 | 85.14 |
| **R-GCN** | **86.36** | 82.08 | 60.78 | 41.56 | 73.95 | 76.18 | **75.00**† | **92.86** | **86.10** | **84.05** | **72.38** | **86.72** | 87.53 | **87.11**† |
| | ±0.31 | ±0.21 | ±0.64 | ±0.86 | ±0.32 | ±0.53 | ±0.18 | ±0.46 | ±0.94 | ±0.74 | ±1.79 | ±0.45 | ±0.34 | ±0.36 |
| **R-GCN (w/o Gate)** | 86.10 | 81.90 | 60.17 | 41.59 | 72.50 | **76.89** | 74.60 | 92.74 | 85.89 | 83.01 | 72.35 | 85.90 | **87.57** | 86.70 |
| | ±0.37 | ±1.74 | ±0.48 | ±0.99 | ±0.79 | ±0.79 | ±0.36 | ±0.86 | ±1.09 | ±1.07 | ±1.72 | ±1.19 | ±0.49 | ±0.52 |

**Table 3: Ablation study over two main components of proposed model (%).**

| Methods | TWITTER-2015 | | | | | | | TWITTER-2017 | | | | | | |
| | Single Type ($F_1$) | | | | Overall | | | Single Type ($F_1$) | | | | Overall | | |
| | PER | LOC | ORG | MISC | Pre | Rec | F1 | PER | LOC | ORG | MISC | Pre | Rec | F1 |
| R-GCN | 86.36 | 82.08 | 60.78 | 41.56 | 73.95 | 76.18 | 75.00 | 92.86 | 86.10 | 84.05 | 72.38 | 86.72 | 87.53 | 87.11 |
| w/o InterRG | 85.52 | 81.16 | 59.30 | 40.74 | 73.42 | 74.79 | 74.05 | 92.63 | 85.32 | 81.55 | 72.29 | 85.34 | 87.07 | 86.17 |
| w/o IntraRG | 85.41 | 81.75 | 60.66 | 40.01 | 73.15 | 75.55 | 74.29 | 92.90 | 82.58 | 82.73 | 71.82 | 85.44 | 87.05 | 86.22 |
| w/o InterRG, IntraRG | 85.51 | 81.45 | 58.72 | 37.61 | 73.18 | 74.09 | 73.59 | 93.58 | 81.59 | 80.12 | 71.37 | 84.12 | 86.21 | 85.13 |

most similar image information into the text sequence contributes more to our model. This is consistent with our motivation.

(3) To better understand the advantage of *IntraRG* and *InterRG*, we shows the qualitative results compared with two state-of-the-art methods. As shown in Figure 4(a), the named entity "KyrieIrving" in the sentence does not appear in the corresponding image, so *UMT* and *UMGF* incorrectly predict the named entity tag as "MISC". However, with the help of *InterRG*, the sentence is able to make the connection with other images in the dataset, thereby giving the right prediction "PER" since these relevant images all contain the visual object *person*. Moreover, in Figure 4(b), the visual object region is fuzzy, which brings the challenge to the recognition of named entity. Both *UMT* and *UMGF* believe that there is no named entity in the sentence. However, with the help of *IntraRG*, we gather similar images involving clear visual object regions into

current image to make the right prediction, since these clear visual object regions reduce the difficulty of identifying named entity.

## 5.3 Discussion

*5.3.1 Effect of hyper-parameter K.* We tune the value of hyper-parameters $K$ on the development set of each dataset, and then evaluate the robustness of the model on the test set. As shown in Table 4, we separately extract the top 1, 3, 5 and 7 image information most relevant to the current image and text from the dataset. Obviously, as the value of $K$ increases, the performance of *R-GCN* gets better, and our model achieves the best results when the value of $K$ is equal to 5. However, once the value of $K$ is greater than 5, the performance does not continue to increase and even begins to fall. A possible reason is that fusing too many image information will bring some noise to our model. Besides, we notice that the best

**Table 4: Effect of hyper-parameter $K$ (%).**

| | TWITTER-2015 | | | | | | TWITTER-2017 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall (Dev) | | | Overall (Test) | | | Overall (Dev) | | | Overall (Test) | | |
| | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** |
| TOP-1 | 74.21 | 72.61 | 73.19 | 72.88 | 75.41 | 74.08 | 86.35 | 86.05 | 86.09 | 87.03 | 85.96 | 86.38 |
| TOP-3 | 73.79 | 73.83 | 73.62 | 75.22 | 75.38 | 74.47 | 86.97 | 86.73 | 86.74 | 86.11 | 87.38 | 86.73 |
| TOP-5 | 73.96 | 73.99 | 73.85 | 73.95 | 76.18 | 75.00 | 87.59 | 86.47 | 86.93 | 86.72 | 87.53 | 87.11 |
| TOP-7 | 73.71 | 73.95 | 73.65 | 73.90 | 75.53 | 74.66 | 87.29 | 86.52 | 86.80 | 86.18 | 87.52 | 86.82 |

**Table 5: Effect of different image encoders(%).**

| | TWITTER-2015 | | | | | | TWITTER-2017 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall (Dev) | | | Overall (Test) | | | Overall (Dev) | | | Overall (Test) | | |
| | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** | **Pre** | **Rec** | **F1** |
| R-GCN | 73.96 | 73.99 | 73.85 | 73.95 | 76.18 | 75.00 | 87.59 | 86.47 | 86.93 | 86.72 | 87.53 | 87.11 |
| R-GCN (Faster-RCNN) | 74.17 | 74.31 | 74.03 | 73.76 | 76.63 | 75.13 | 87.99 | 86.26 | 87.03 | 86.79 | 87.69 | 87.22 |
| R-GCN (Mask-RCNN) | 74.24 | 74.70 | 74.24 | 74.32 | 76.11 | 75.16 | 87.26 | 87.06 | 87.10 | 86.73 | 87.85 | 87.28 |
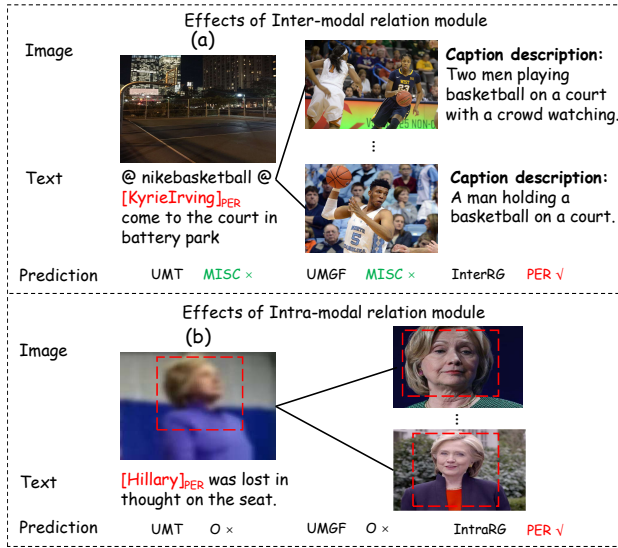


**Figure 4: Predictions of *UMT*, *UMGF*, *InterRG* and *IntraRG* on two test samples. ✗ and ✔ denote incorrect and correct predictions. The named entities and their corresponding types are highlighted in the text. "O" means not a named entity.**

results of the development set and test set are basically consistent, indicating the robustness of our model.

*5.3.2 Effect of Different Image Encoder.* In order to explore the impact of different image encoders on our *R-GCN* model, we conduct experiments (shown in Table 5) using the pre-trained image recognition models Faster-RCNN [21] and Mask-RCNN [9]. Specifically, we replace the ResNet encoder with Faster-RCNN or Mask-RCNN and keep the other components the same as our original model. It's clear that the performance of *R-GCN*(Faster-RCNN) and

*R-GCN*(Mask-RCNN) obtain the improvement since we use a more powerful image recognition encoder. However, *R-GCN*(Faster-RCNN) and *R-GCN*(Mask-RCNN) perform slightly better than the model *R-GCN*, it's also reasonable since our propose inter-modal relation module and intra-modal relation module are effective enough, so our overall performance is less affected by the image encoder.

*5.3.3 Time Complexity.* The time complexity of *R-GCN* is $O((2K + 3)nd^2 + 3n^2d)$, while the time complexity of latest multimodal NER models *RpBERT*, *UMT*, and *UMGF* is $O(Lnd^2 + Ln^2d)$, where $K$ and $L$ are constants ($<= 10$ usually). Hence, in terms of time complexity, we remain in the same order of magnitude as other related works.

Besides, due to space constraints, we present the **Error Analysis** in the Appendix.

## 6 CONCLUSION

In this paper, we propose a novel Relation-enhanced Graph Convolutional Network (*R-GCN*) for the Multimodal Named Entity Recognition (MNER) task. The main idea of our approach is to leverage two kinds of external matching relations (i.e., inter-modal and intra-modal relations) in different (image, text) pairs to improve the ability of identifying named entities in the text. Results from numerous experiments indicate that our model achieves better performance than other state-of-the-art methods. Further analysis also validates the effectiveness of *R-GCN* model.

In the future, we would like to apply our idea to other multimodal tasks since the external matching relations between different (text, image) pairs is easy to extend to other multimodal tasks, such as multimodal dialogue or multimodal entailment.

## ACKNOWLEDGMENTS

# REFERENCES

[1] Md. Shad Akhtar, Tarun Garg, and Asif Ekbal. 2020. Multi-task learning for aspect term extraction and aspect sentiment classification. *Neurocomputing* 398 (2020), 247–256.

[2] Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal Named Entity Recognition with Image Attributes and Image Knowledge. In *DASFAA 2021, Taipei, Taiwan, April 11-14, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12682)*. Springer, 186–201.

[3] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Can images help recognize entities? A study of the role of images for Multimodal NER. In *W-NUT 2021, Online, November 11, 2021*. Association for Computational Linguistics, 87–96.

[4] Jason P. C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguistics* 4 (2016), 357–370.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.

[6] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. 2021. Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning. In *MM '21, Virtual Event, China, October 20 - 24, 2021*. ACM, 2615–2624.

[7] Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 2619–2629.

[8] Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks*, Vol. 2. 729–734.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2020. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2 (2020), 386–397.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778.

[11] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580 (2012). arXiv:1207.0580

[12] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015). arXiv:1508.01991

[13] Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *EMNLP-CoNLL 2007, June 28-30, 2007, Prague, Czech Republic*. ACL, 698–707.

[14] Zaid Khan and Yun Fu. 2021. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. In *MM '21, Virtual Event, China, October 20 - 24, 2021*. ACM, 3034–3042.

[15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[16] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

[17] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *NAACL HLT 2016, San Diego California, USA, June 12-17, 2016*. The Association for Computational Linguistics, 260–270.

[18] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. In *ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Association for Computational Linguistics, 1990–1999.

[19] Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

[20] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. In *NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Association for Computational Linguistics, 852–860.

[21] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. 91–99.

[22] Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing Text Chunks. In *EACL 1999, June 8-12, 1999, University of Bergen, Bergen, Norway*. The Association for Computer Linguistics, 173–179.

[23] Chengai Sun, Liangyu Lv, Gang Tian, and Tailu Liu. 2021. Deep Interactive Memory Network for Aspect-Level Sentiment Analysis. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 20, 1 (2021), 3:1–3:12.

[24] Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A Pre-trained Tweet Multimodal Model Based on Text-image Relation for Multimodal NER. In *COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*. International Committee on Computational Linguistics, 1852–1862.

[25] Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rp-BERT: A Text-image Relation Propagation-based BERT Model for Multimodal NER. In *AAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 13860–13868.

[26] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Un-aligned Multimodal Language Sequences. In *ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, 6558–6569.

[27] Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly Simple Unsupervised Aspect Extraction. In *ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 3182–3187.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008.

[29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[30] Hanqian Wu, Siliang Cheng, Jingjing Wang, Shoushan Li, and Lian Chi. 2020. Multimodal Aspect Extraction with Region-Aware Alignment Network. In *NLPCC 2020, Zhengzhou, China, October 14-18, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12430)*. Springer, 145–156.

[31] Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Multimodal Representation with Embedded Visual Guiding Objects for Named Entity Recognition in Social Media Posts. In *MM '20, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. ACM, 1038–1046.

[32] Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design Challenges and Misconceptions in Neural Sequence Labeling. In *COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. Association for Computational Linguistics, 3879–3889.

[33] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, 3342–3352.

[34] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel Methods for Relation Extraction. In *EMNLP 2002, Philadelphia, PA, USA, July 6-7, 2002*. 71–78.

[35] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing. In *AAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 14338–14346.

[36] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance. In *AAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 14347–14355.

[37] Min Zhang, Guodong Zhou, Lingpeng Yang, and Dong-Hong Ji. 2006. Chinese Word Segmentation and Named Entity Recognition Based on a Context-Dependent Mutual Information Independence Model. In *SIGHAN@COLING/ACL 2006, Sydney, Australia, July 22-23, 2006*. Association for Computational Linguistics, 154–157.

[38] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In *(AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 5674–5681.

[39] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-Aware BERT for Language Understanding. In *AAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 9628–9635.

[40] Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2021. Object-Aware Multimodal Named Entity Recognition in Social Media Posts With Adversarial Learning. *IEEE Trans. Multim.* 23 (2021), 2520–2532.

[41] Guodong Zhou and Jian Su. 2005. Machine learning-based named entity recognition via effective integration of various evidences. *Nat. Lang. Eng.* (2005).

# A FULL RESULTS WITH STANDARD DEVIATIONS

Here, we present the results with standard deviations for ablation study in Table 6.

**Table 6: Ablation study over two main components of proposed model (%).**

| Methods | TWITTER-2015 | | | | | | | TWITTER-2017 | | | | | | |
| | Single Type ($F_1$) | | | | Overall | | | Single Type ($F_1$) | | | | Overall | | |
| | PER | LOC | ORG | MISC | Pre | Rec | F1 | PER | LOC | ORG | MISC | Pre | Rec | F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| R-GCN | 86.36 | 82.08 | 60.78 | 41.56 | 73.95 | 76.18 | 75.00 | 92.86 | 86.10 | 84.05 | 72.38 | 86.72 | 87.53 | 87.11 |
| | ±0.31 | ±0.21 | ±0.64 | ±0.86 | ±0.32 | ±0.53 | ±0.18 | ±0.46 | ±0.94 | ±0.74 | ±1.79 | ±0.45 | ±0.34 | ±0.36 |
| w/o InterRG | 85.52 | 81.16 | 59.30 | 40.74 | 73.42 | 74.79 | 74.05 | 92.63 | 85.32 | 81.55 | 72.29 | 85.34 | 87.07 | 86.17 |
| | ±0.43 | ±0.41 | ±0.30 | ±0.19 | ±0.36 | ±0.38 | ±0.37 | ±0.46 | ±0.41 | ±0.41 | ±0.36 | ±0.43 | ±0.44 | ±0.43 |
| w/o IntraRG | 85.41 | 81.75 | 60.66 | 40.01 | 73.15 | 75.55 | 74.29 | 92.90 | 82.58 | 82.73 | 71.82 | 85.44 | 87.05 | 86.22 |
| | ±0.49 | ±0.47 | ±0.34 | ±0.22 | ±0.42 | ±0.43 | ±0.43 | ±0.53 | ±0.48 | ±0.47 | ±0.42 | ±0.49 | ±0.50 | ±0.50 |
| w/o InterRG, IntraRG | 85.51 | 81.45 | 58.72 | 37.61 | 73.18 | 74.09 | 73.59 | 93.58 | 81.59 | 80.12 | 71.37 | 84.12 | 86.21 | 85.13 |
| | ±0.35 | ±0.47 | ±1.02 | ±0.85 | ±0.52 | ±0.22 | ±0.31 | ±0.40 | ±0.79 | ±0.49 | ±1.16 | ±0.45 | ±0.38 | ±0.26 |

## B ERROR ANALYSIS

We randomly sample 100 error cases of R-GCN model, and classify them into three error categories. Figure 5 shows the proportions and some representative examples for each category. The top category is bias brought by annotation. As shown in Figure 5(a), the named entity "Pebble Beach Residence" is annotated as "ORG", but it is also reasonable if we annotate it as an "LOC". In this case, it's really challenge for our model to distinguish them since they are all correct. The second category is lack of background knowledge. In Figure 5(b), the entity "Jonas brother" is the name of a famous band, we are easily to misunderstand this entity as "PER" without the help of background knowledge. The third category is information deficiency. As shown in Figure 5(c), the sentence is very short and the content of the image is also very simple, which is unable to provide sufficient information for our model to tell the entity

type. There ought to be more advanced natural language processing techniques developed to address them.



**Figure 5: Three typical errors of R-GCN.**