

# P-MNER: Cross Modal Correction Fusion Network with Prompt Learning for Multimodal Named Entity Recognition

Zhuang Wang<sup>1</sup>, Yijia Zhang<sup>1\*</sup>, Kang An<sup>1</sup>, Xiaoying Zhou<sup>1</sup>, Mingyu Lu<sup>2\*</sup>, Hongfei Lin<sup>3</sup>

<sup>1</sup>College of Information Science and Technology, Dalian Maritime University / Dalian, China

<sup>2</sup>College of Artificial Intelligence, Dalian Maritime University / Dalian, China

<sup>3</sup>College of Computer Science and Technology, Dalian University of Technology / Dalian, China

{wang\_1120211498,zhangyijia,1120221416\_ankang,zhouxiaoying}@dlmu.edu.cn

lumingyu@dlmu.edu.cn

hflin@dlut.edu.cn

## Abstract

Multimodal Named Entity Recognition (MNER) is a challenging task in social media due to the combination of text and image features. Previous MNER work has focused on predicting entity information after fusing visual and text features. However, pre-training language models have already acquired vast amounts of knowledge during their pre-training process. To leverage this knowledge, we propose a prompt network for MNER tasks (P-MNER). To minimize the noise generated by irrelevant areas in the image, we design a visual feature extraction model (FRR) based on FasterRCNN and ResNet, which uses fine-grained visual features to assist MNER tasks. Moreover, we introduce a text correction fusion module (TCFM) into the model to address visual bias during modal fusion. We employ the idea of a residual network to modify the fused features using the original text features. Our experiments on two benchmark datasets demonstrate that our proposed model outperforms existing MNER methods. P-MNER's ability to leverage pre-training knowledge from language models, incorporate fine-grained visual features, and correct for visual bias, makes it a promising approach for multimodal named entity recognition in social media posts.

## 1 Introduction

With the rapid development of the Internet, social media platforms have experienced an exponential growth of content. These platforms offer a wealth of user-generated posts that provide valuable insights into the events, opinions, and preferences of both individuals and groups. Named Entity Recognition (NER) is a crucial task in which entities contained in textual data are detected and mapped to predefined entity types, such as location (LOC), person (PER), organization (ORG), and miscellaneous (MISC). Incorporating visual information from posts has been shown to significantly enhance the accuracy of entity prediction from social media content. For instance, as illustrated in Fig.1, the sentence "Alban got Rikard a snowball in the snow" can be easily resolved by leveraging the visual cues in the accompanying image, allowing us to identify "Rikard" as an animal. However, relying solely on textual data to predict entities may lead to erroneous predictions, such as identifying "Rikard" as a name.

With the continuous evolution of deep learning models, several multi-modal Named Entity Recognition (NER) models have been proposed to enhance the prediction performance of entities by incorporating visual information. These models employ techniques such as cross-attention(Wu et al., 2020; Zhang et al., 2018), adversarial learning(Goodfellow et al., 2014; Frankle and Carbin, 2018), and graph fusion(Xiao et al., 2021; Wu et al., 2020). However, previous methods fused text features with visual features and directly fed them into a neural network model for prediction. This approach overlooks the wealth of information embedded

---

This work is supported by the National Natural Science Foundation of China (No.61976124) ©2023 China National Conference on Computational Linguistics

Published under Creative Commons Attribution 4.0 International License



Figure 1: An example for MNER with (A and B) the useful visual clues and the (C and D) useless visual clues.

in the pre-training language model itself. To overcome this limitation, we propose the use of prompt learning (Liu et al., 2023) to process the fused features, followed by final training.

The presence of irrelevant content in an image may negatively impact the performance of Named Entity Recognition (NER) models. As illustrated in Fig.1, regions A and B in an image may aid in identifying entities in a sentence, while regions C and D may not contribute to model prediction. In previous Multi-modal NER (MNER) tasks, however, all visual regions were involved in cross-modal fusion. To address this issue, we propose a novel model (FRR), which utilizes visual objects in images for modal fusion. This approach effectively eliminates extraneous image features that are irrelevant to the corresponding text.

In this paper, we present a new Transformer-based (Vaswani et al., 2017) text correction fusion module (TCFM) to address the issue of cross-modal visual bias in the named entity recognition (NER) task. Inspired by the residual network, the TCFM continuously integrates the original text features with the fusion features to iteratively correct the fusion features. This approach effectively alleviates the problem of visual bias and enhances the performance of the NER task.

In order to showcase the effectiveness of our proposed approach, we conducted a comprehensive set of experiments on two publicly available datasets: Twitter-2015 and Twitter-2017. The obtained experimental results unequivocally demonstrate that our method outperforms the existing MNER algorithm in terms of performance.

The significant contributions of our work can be summarized as follows:

- We introduce a novel approach, the Prompt Network for Named Entity Recognition (P-MNER), which aims to leverage the abundant information present in pre-trained language models. To accommodate the specific requirements of our proposed prompt network, we further present a novel Text Correction Fusion Module (TCFM) that effectively minimizes the visual bias in the fusion process.
- To mitigate the impact of irrelevant visual regions on modal fusion, we propose a novel Feature Extraction Module (FRR) that leverages fine-grained visual objects for more precise feature extraction.
- Experimental results show that our proposed P-MNER network achieves SOTA performance on both datasets.

## 2 Related work

Named Entity Recognition (NER) has emerged as a crucial component in a plethora of downstream natural language processing (NLP) applications, including but not limited to affective

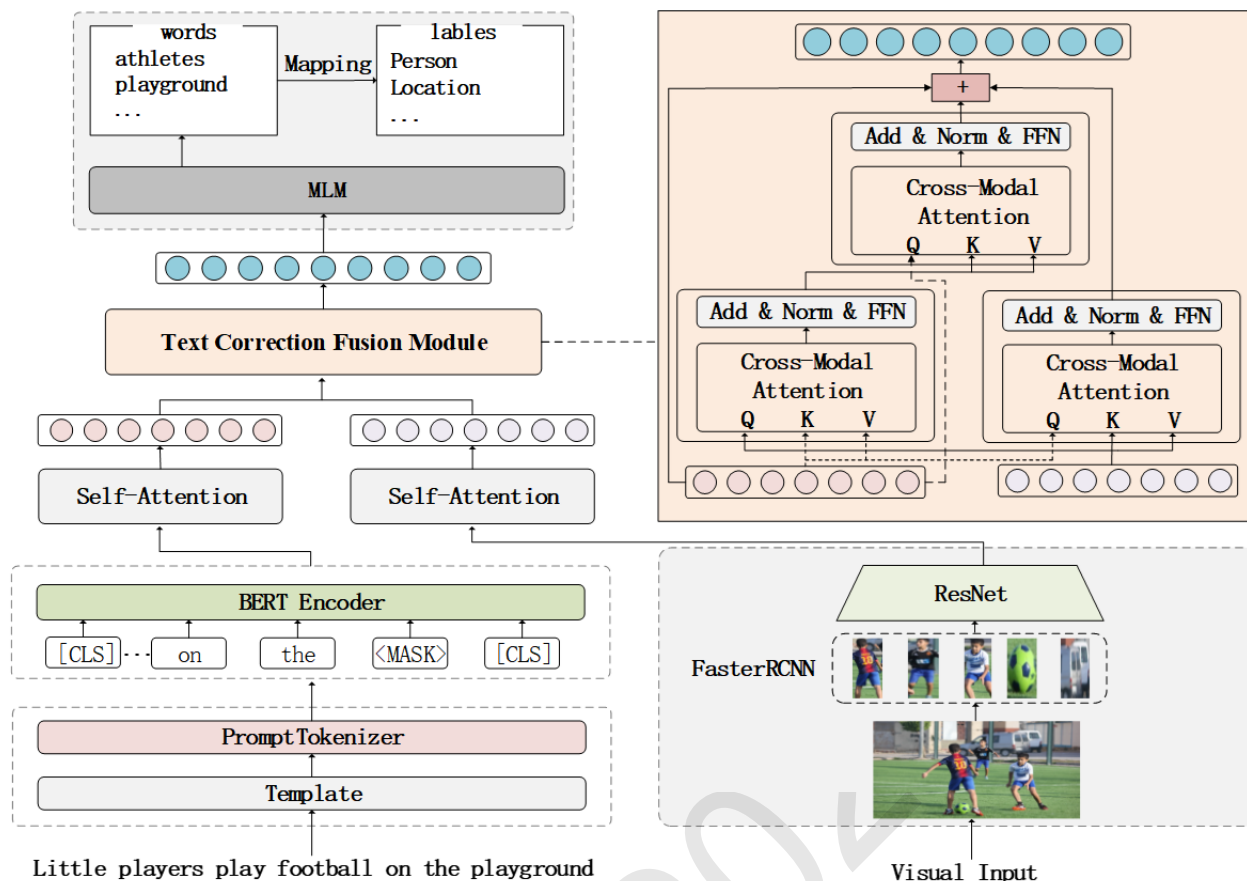


Figure 2: The overall architecture of our P-MNER model.

analysis(Chen et al., 2022), relationship extraction(Gupta et al., 2017), and knowledge graph(Cui et al., 2021) construction. With the advent of neural network models, the use of Bi-LSTM(Chiu and Nichols, 2016) or Convolutional Neural Network (CNN)(Zhao et al., 2017) as encoders, and Softmax(Joulin et al., 2017), RNN(Liu et al., 2016), or Conditional Random Field (CRF)(Zhuo et al., 2016) as decoders has gained popularity in the NER community. For instance, Huang et al.(Huang et al., 2015) utilized BiLSTM and CRF as the encoder and decoder, respectively, for NER tasks. Similarly, Chiu and Nichols et al.(Chiu and Nichols, 2016) proposed a CNN-based encoder with CRF as the decoder to accomplish the final prediction.

Social media posts are characterized by their brevity and high levels of noise, which often lead to suboptimal performance of conventional NER methods when applied to such data. To address this challenge, several recent studies have proposed novel approaches for cross-modal fusion in the context of NER. For instance, Sun et al.(Ritter et al., 2011; Sun et al., 2021) introduced a novel image-text fusion approach for NER tasks, while Zhang et al.(Zhang et al., 2018) employed BiLSTM to combine visual and textual features in multimodal social posts. Similarly, Zheng et al.(Zheng et al., 2020) proposed an adversarial learning approach to tackle the issue of semantic gap in multimodal NER tasks. Lu et al.(Lu et al., 2018) integrated an attention mechanism into the modal fusion process and introduced a visual gate to filter out noise in the image. Lastly, Yu et al.(Yu et al., 2020) designed a multimodal Transformer model for the MNER task and introduced an entity span module to facilitate the final prediction.

In order to utilize the vast amount of knowledge encoded in pre-trained language models, Wang et al.(Wang et al., 2022) proposed a prompt-based method, namely PromptMNER, which extracts visual features and subsequently fuses them with input text for enhanced performance.

Previous studies on named entity recognition (NER) in social media have yielded promising results. However, these methods have been unable to fully utilize the power of pre-training language models in capturing the contextual nuances of input text. This shortcoming has limited the overall effectiveness of NER in social media. To overcome this challenge, our paper proposes a novel prompt learning approach that directly leverages the knowledge embedded in pre-training language models to enrich input text and image features. By doing so, we are able to tap into the full potential of these models and achieve a higher level of accuracy in NER tasks. In essence, our approach represents a significant step forward in the field of NER for social media. It addresses a critical limitation of previous methods and provides a more effective way of leveraging pre-training language models. Our approach offers a promising new avenue for future research.

### 3 Methods

The objective of this study is to predict a tag sequence  $L = (l_1, l_2, \dots, l_n)$  given a sentence  $Y$  and an associated picture  $V$  as input. Here,  $l_i$  belongs to a predefined set of tags for the BIOES tagging pattern.

Fig.2 depicts the overall architecture of our proposed model, which comprises four modules: visual feature extraction, text feature extraction, modal fusion, and prompt learning. The visual feature extraction module takes the objects in the picture as input. In the text feature extraction module, we leverage BERT to extract features by processing the wrapped text. We also introduce a text correction fusion module to obtain more precise fusion features. In the prompt learning module, we utilize the Bert Masked Language Model for prompt learning.

#### 3.1 Text Feature Extraction

In Fig.2, the input sentence  $Y = (y_1, y_2, \dots, y_n)$  is demonstrated on the left. To structure the sentences, a wrapping class is utilized, which adheres to a predetermined template. For MNER direction, a prompt template is introduced: “<sentence>, the word of <entity> is <mask>.” Within the template, <sentence> represents the sentence  $Y$ , <entity>  $\in \mathcal{Y}$ .

In order to effectively process complex information from input text and templates, a new Tokenizer has been introduced to tokenize the input sentences that are wrapped by the wrapper class. Following the pre-training models, special tokens are added to the tokenized sequences to form a sequence, denoted as  $S = (s_0, s_1, \dots, s_{n+1})$ , where  $s_0$  and  $s_{n+1}$  represent the two special tokens at the beginning and end of the final sequence. The tokenized sequences are then sent to the embedding layer where BERT embedding is utilized to convert each word into a vector form that includes token embedding, segment embeddings, and position embeddings.

$$x_i = e^t(s_i) + e^s(s_i) + e^p(s_i) \quad (1)$$

where  $\{e^t, e^s, e^p\}$  denotes the embeddings lookup table.  $X = (x_0, x_1, \dots, x_{n+1})$  is the word representation of  $S$ , where  $x_i$  is the sum of word, segment, and position embeddings for token  $y_i$ .

In various social media posts, the same word may have different meanings depending on the context. To address this challenge, we adopt BERT as the sentence encoder. The resulting embedding representation is fed into the BERT encoder, producing a signature of encodings  $R = \{r_0, r_1, \dots, r_{n+1}\}$ .

The Self-Attention mechanism establishes direct links between any two words in a sentence through a calculation step, significantly reducing the distance between distance-dependent features and enabling their efficient use. Consequently, we feed the hidden representation output by BERT Encoder into Self-Attention to capture long-distance dependencies in a sentence.

$$T = \text{softmax} \left( \frac{[W_{qt}R]^T [W_{kt}RR]}{\sqrt{d_t}} \right) [W_{vt}R]^T \quad (2)$$

where  $\{W_{k_i}, W_{v_i}, W_{q_i}\}$  is parameter matrices for the key, value and query. The final text feature  $T = \{t_0, t_1, \dots, t_{n+1}\}$ , where  $t_i$  is the generated contextualized representation for  $y_i$ .

### 3.2 Visual Feature Extraction

We utilize object-level visual features in the visual feature extraction module to aid named entity recognition, and introduce a novel method for feature extraction.

To begin with, we feed the image into the Faster RCNN (Faster, 2015) detection module to extract the visual object area. Specifically, we input the image into a feature extraction network that includes convolutional layers, pooling layers, and rectified linear unit (ReLU) layers to obtain feature maps of the image. Next, we pass the feature maps to the Region Proposal Networks (RPN) to train them to extract Region Proposal regions from the original maps. Then, we use RoI Pooling to normalize candidate recognition areas of different sizes and shapes into fixed-size target recognition areas. RoI Pooling collects proposals (coordinates of each box) generated by RPN and extracts them from feature maps. Finally, we process the resulting proposals with Fully Connected and Softmax to determine the probability that each proposal corresponds to a particular category.

Typically, only a small number of visual entities are needed to emphasize the entities in a sentence. To accomplish this, we choose the first  $m$  visual objects with a probability exceeding 0.95. Then, we crop the original picture based on these proposals to obtain the final visual object set, denoted as  $I = \{I_1, I_2, \dots, I_m\}$ , where  $I_i$  represents the  $i$ -th visual object.

The residual network is among the most advanced CNN image recognition models, with the ability to extract meaningful features from input images. Thus, we feed the resulting visual objects into a pre-trained 152-layer ResNet and use the output of the last convolution layer as the visual characteristics of each object, denoted as  $\tilde{V} = \{\tilde{V}_1, \tilde{V}_2, \dots, \tilde{V}_m\}$ , where  $\tilde{V}_i \in R^{1024}$  represents the features of the  $i$ -th object. We then employ Self-Attention to enable each visual block to fully comprehend the context of the visual features:

$$V = \text{softmax} \left( \frac{[W_{q_v} \tilde{V}]^T [W_{k_v} \tilde{V}]}{\sqrt{d_v}} \right) [W_{v_v} \tilde{V}]^T \quad (3)$$

where  $\{W_{q_v}, W_{k_v}, W_{v_v}\}$  denote the weight matrices for the query, key and value. The final visual features are:  $V = \{v_1, v_2, \dots, v_m\}$ ,  $v_i$  refer to the visual features processed by Self-Attention.

### 3.3 Text Correction Fusion Module

In the text feature extraction module, we have extracted text features through con-textual comprehension. However, the short length of social media posts and the presence of irrelevant information make it challenging to accurately identify entities using text information alone. To address this issue, we utilize visual objects in pictures to guide text-based word representations for improved accuracy. Nevertheless, the challenge of visual bias in modal fusion remains. Therefore, we propose a text correction fusion module to generate the final fusion features.

As shown in the right of Fig.2, we initially apply a  $k$ -head cross-modal attention mechanism. This involves using the visual features  $V = \{v_1, v_2, \dots, v_m\}$  as queries in the self-attention mechanism and utilizing the text features  $T = \{t_0, t_1, \dots, t_{n+1}\}$  as keys and values:

$$H_i(V, T) = \text{softmax} \left( \frac{[W_{q_i} V]^T [W_{k_i} T]}{\sqrt{d/k}} \right) [W_{v_i} T]^T \quad (4)$$

$$M-H(V, T) = W' [H_1(V, T), \dots, H_k(V, T)]^T \quad (5)$$

where  $H_i$  refers to the  $i$ -th head of cross-modal attention,  $\{W_{q_i}, W_{k_i}, W_{v_i}\}$  and  $W'$  denote the weight matrices for the query, key, value, and multi-head attention, respectively. By utilizing

this cross-attention approach, we can derive feature representations based on the correlation between words and visual objects in the text. We then process the fused features through two normalization layers and a feed-forward neural network (Vaswani et al., 2017):

$$\tilde{P} = LN(V + M_{-}H(V, T)) \quad (6)$$

$$P = LN(\tilde{P} + FFN(\tilde{P})) \quad (7)$$

where FFN is the feed-forward network, LN is the layer normalization. Get the text features based on visual objects, denoting as  $P = \{p_0, p_1, \dots, p_{n+1}\}$ . Similar to the description above, We use the text feature  $T = \{t_0, t_1, \dots, t_{n+1}\}$  as queries in our own attention and the visual feature  $V = \{v_1, v_2, \dots, v_m\}$  as keys and values. The result is a text-based visual object, denoting as  $q = \{q_1, q_2, \dots, q_m\}$ .

During the process of acquiring visual object-based text features, the resulting features may exhibit bias towards the visual mode, as the queries used are primarily based on visual features. In order to alleviate such bias, we propose the use of a cross-modal layer for the re-fusion of text features. In this approach, the original text features are employed as queries, while the visual-based text fusion features are utilized as keys and values. The final cross-modal text representation is obtained as  $C = \{c_0, c_1, \dots, c_{n+1}\}$ .

Previous studies have simply connected cross-modal visual features and cross-modal text features, which may lead to biased final fusion features. In this paper, we propose an alternative approach for the final stitching process by connecting initial text features to both cross-modal visual features and cross-modal text features. This method aims to mitigate bias and enhance the quality of the fusion features.

$$H = T + V + C \quad (8)$$

where T is the initial text features, V is the cross-modal visual features, and C is the cross-modal text features.

By incorporating the original text features in the final fusion process, it is effectively reduce visual bias. The resulting fusion feature is denoted as  $H = \{h_0, h_1, \dots, h_{n+1}\}$ .

### 3.4 Prompt-Learning Module

In this module, we employ the Bert model as our Pre-trained Language Model (PLM). Our approach involves inputting the resulting fusion feature H into the PLM and leveraging the masked language model (MLM) to reconstruct sequences with <MASK>. The predicted part of the text is replaced with <MASK> during packaging to optimize the pre-training language model stimulation. Our method follows the pre-training language model training process for processing fusion features.

In PLM, our aim is to predict a probability distribution for the <MASK> section that aligns with the objectives of MLM. Here, we are only predicting that part of <MASK> belongs to a certain vocabulary. The ultimate goal is to predict <MASK> as predefined tags in a sentence. To accomplish this, we introduce a verbalizer class to process the output of the MLM model. This class constructs a mapping from original tags to words. When PLM predicts a probability distribution for a masked location in the vocabulary, the verbalizer maps the word to the original label. The output layer can be defined as:

$$c_i = \text{plm}(h_i) \quad (9)$$

$$d_i = \text{ver}(c_i) \quad (10)$$

Table 1: Statistics of Twitter datasets.

Entity Type	Train-15	Dev-15	Test-15	Train-17	Dev-17	Test-17
PER	2217	552	1816	2943	626	621
LOC	2091	552	1697	731	173	178
ORG	928	247	839	1674	375	395
MISC	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
Tweets	4000	1000	3257	3373	723	723

where  $plm$  is masked language model (MLM),  $ver$  refers to the verbalizer.  $c_i$  is the probability distribution of predicted positions on the vocabulary,  $d_i$  is a label for prediction. Finally, the prediction tag distribution is  $D = \{d_0, d_1, \dots, d_{n+1}\}$ .

During the training phase, we calculate the loss of verbalizer-mapped labels and real labels:

$$L = - \sum_{i=1}^n o_i \log(d_i) \quad (11)$$

where  $o_i$  is the true tag for  $d_i$ .

## 4 Experiments

We tested the model on two common datasets. Furthermore, we compare our model with the single-mode NER model and the existing multimodal methods.

### 4.1 Experiment Settings

**Datasets:** During the model training and evaluation phase, we employed a publicly available dataset from Twitter, comprising four distinct entity types, namely PER, LOC, ORG, and MISC, with non-entity words marked as O. Following the same protocol established by Zhang et al. (Zhang et al., 2018), the dataset was partitioned into training, development, and test sets. Table 1 provides an overview of the dataset, including the number of samples in each set and the count of each entity type.

**Hyperparameter:** Compared with other NER methods, our model is an experiment performed on a GUP. For visual object extraction, the first five objects with an accuracy above 0.95 are selected for feature extraction using a pre-trained 152-layer Res-Net. The maximum sentence length is set to 128, and the batch size is 8. The input template has a maximum length of 20, while the encoded text length is set at 256. Cross-modal multi-head attention is applied to facilitate modal fusion, utilizing 12 attention heads. The learning rate and learning attenuation rate are set at 0.005 and 0.01, respectively. During the evaluation phase, standard precision, recall rate, and F1-score are employed as evaluation metrics. The model with the highest performance in the evaluation phase is selected, and its performance is reported on the test dataset.

### 4.2 Main Result

Table 2 presents the experimental results of our proposed model and the comparative approaches. During model evaluation, we calculated the precision (P), recall (R), and F1-score (F1) of our model.

In the upper section of Table 2, we initially conducted a series of experiments using a text-only model to extract features. Our findings revealed that employing BERT as the encoder for text feature extraction resulted in significantly superior results compared to other methods. We believe that the contextualized word representation and contextual understanding of the input

Table 2: Performance comparison on two TWITTER datasets. Specifically, B-L+CRF and C+B-L+CRF refers to Bi-LSTM+CRF and CNN+Bi-LSTM+CRF, respectively.

Models	TWITTER-2015			TWITTER-2017		
	P	R	F1	P	R	F1
B-L+CRF	68.14	61.09	64.42	79.42	73.42	76.31
C+B-L+CRF	66.24	68.09	67.15	80.00	78.76	79.31
T-NER	69.54	68.65	69.09	-	-	-
BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
MNER-MA	72.33	63.51	67.63	-	-	-
AGBAN	74.13	72.39	73.25	-	-	-
UMT	71.67	75.23	73.41	85.28	85.34	85.31
UMGF	74.49	75.21	74.85	86.54	84.50	85.51
PromptMNER	78.03	79.17	78.60	89.93	90.10	90.27
Ours	79.18	79.55	79.43	90.11	91.23	91.31

text played a crucial role in enhancing the performance of the NER models. In order to achieve even deeper text representation, we leveraged BERT to extract hidden features of the text.

Moreover, we took our analysis a step further and experimented with some representative multimodal NER models to compare their performance with single-mode NER models. As shown in Table 2, the results demonstrate that MNER-MA outperforms the single-mode NER models, indicating the effectiveness of combining visual information in NER tasks. However, we noticed that when BERT was utilized to replace the encoder in the model, the observed improvement was relatively modest. Therefore, it is evident that novel methods need to be developed and employed to address the current limitations in this area.

Prompt learning, a novel paradigm, has demonstrated strong potential in the field of NLP. Wang et al. (Wang et al., 2022) propose utilizing prompt learning to aid in the extraction of visual features. Specifically, they suggest employing the CLIP model as a prompt language model (PLM) to leverage the learned information from the pre-training stage for visual feature extraction. During training, both visual and text information are processed and fed into the PLM to obtain visual features based on prompts. Finally, the extracted visual and text features are fused together.

The results presented in Table 2 unequivocally demonstrate the superior performance of our proposed approach over existing single-mode approaches in the task of named entity recognition (NER). Our method outperforms the current state-of-the-art MNER method as well, owing to our incorporation of prompt learning, which allows us to extract rich information from the pre-trained language model. The primary reason for our success is the utilization of visual context, which enables us to make full use of the available information and improve the overall accuracy of the model. Our approach outperforms the promptMNER method as well. The incorporation of prompt learning in our model allows us to effectively fuse the visual and text features, thereby making the most of the pre-trained model’s knowledge during the training process. As a result, we are able to achieve a better overall performance in the NER task. In summary, our proposed approach offers a significant improvement over existing single-mode approaches in the NER task. Our method outperforms both the current state-of-the-art MNER method and the promptMNER method. By incorporating visual context and prompt learning, we are able to effectively extract and utilize the rich information contained in the pre-trained language model, resulting in superior performance.



Table 3: The effect of each module in our model.

Models	TWITTER-2015			TWITTER-2017		
T+V	72.76	72.53	72.31	83.74	83.24	84.33
T+V+TCFM	74.38	74.12	73.35	85.16	84.35	85.41
T+FRR+TCFM	75.32	75.54	75.14	85.47	84.89	86.02
OURS	79.18	79.55	79.43	90.11	91.23	91.31

### 4.3 Ablation Result

To evaluate the effectiveness of each component in our proposed P-MNER model, we conduct ablation experiments. Our results, presented in Table 3, indicate that all components in the P-MNER model have contributed significantly to the final predicted results.

T+V is the baseline of our MNER task, with BERT utilized as the encoder for text feature extraction and ResNet employed as the encoder for visual data. The experimental results presented in Table 3 demonstrate that our proposed baseline model achieves a higher F1-score than all single-mode models, thereby validating the effectiveness of incorporating visual information into our model.

T+V+TCFM replaced the modal splicing part with TCFM. Table 3 shows a significant increase in F1-score of 1.22% and 1.08%, respectively, upon implementation of the proposed text correction fusion module, which validates our proposed modal fusion mechanism. Our TCFM module improved accuracy by 1.62% and 1.42% on the two datasets, due to its ability to continuously utilize text information to correct feature bias during mode fusion. This effectively addresses the problem of feature alignment and improves model performance.

T+FRR+TCFM uses a new visual feature extraction method (FRR). Table 3 illustrates that our proposed visual feature extraction module achieved F1-scores of 75.14% and 86.02% on the two datasets, respectively, surpassing other NER methods.

Our proposed model, OURS, is a comprehensive approach that employs prompt learning throughout the entire system. The effectiveness of prompt learning is demonstrated in Table 3, where F1-scores of 79.43% and 91.31% were achieved on two different datasets, respectively, surpassing the current state-of-the-art methods in MNER. The superiority of OURS can be attributed to its ability to deeply explore latent knowledge within pre-trained language models, thanks to the prompt learning technique. Moreover, we achieved a 0.83% and 1.04% improvement in F1-score compared to promptMNER, due to our use of prompt learning for feature fusion processing. Our method is more effective in extracting hidden knowledge from pre-trained language models.

### 4.4 Case analysis

To further strengthen our argument regarding the effectiveness of our proposed method, we have conducted a comprehensive case study analysis. We present the results of this analysis in Fig.3, where we compare the performance of three models for entity prediction: BERT-CRF, UMGF, and P-MNER.

BERT-CRF is a text-only NER model, while UMGF and P-MNER are MNER models that incorporate both visual and textual information. In the first case of our analysis, BERT-CRF failed to accurately predict the entity "Susie". We attribute this to the model's lack of attention to visual information. This highlights the importance of incorporating visual data to improve entity prediction accuracy.

In the second case, all three models correctly predicted the entities. However, this case also revealed that not all image information is semantically consistent with the accompanying text. Hence, the incorporation of visual data should be done thoughtfully and with a proper




<b>Visual Modality</b>			
<b>Textual Modality</b>	[Susie MISC] is playing football in the [park LOC]	[Sonam Kapoor PER] to walk on red carpet at [cannes film festival MISC]	[NFT ORG] star [patrick willis PER] is thriving in retirement as a [silicon valley LOC] tech worker
<b>BERT-CRF</b>	(Susie MISC) (park LOC)	× (Sonam Kapoor PER) ✓ (cannes film festival MISC)	✓ (NFT ORG) ✓ (patrick willis PER) × (silicon valley LOC)
<b>UMGF</b>	(Susie MISC) (park LOC)	✓ (Sonam Kapoor PER) ✓ (cannes film festival MISC)	✓ (NFT ORG) ✓ (patrick willis PER) × (silicon valley LOC)
<b>P-MNER</b>	(Susie MISC) (park LOC)	✓ (Sonam Kapoor PER) ✓ (cannes film festival MISC)	✓ (NFT ORG) ✓ (patrick willis PER) ✓ (silicon valley LOC)

Figure 3: Three cases of the predictions by BERT-CRF, UMGF and OUR MODE

understanding of the context.

Finally, in the third case, both BERT-CRF and UMGF failed to accurately predict the entity types. In contrast, our P-MNER model leverages the pre-trained language model to effectively acquire knowledge and make accurate entity type predictions. Our model outperformed the other models by a considerable margin, thereby highlighting the superiority of our proposed method.

In conclusion, the case study analysis provides strong evidence to support our claim that incorporating visual information enhances the accuracy of entity prediction. Additionally, our proposed P-MNER model outperforms the other models by leveraging the pre-trained language model to acquire knowledge and make accurate predictions.

## 5 Conclusion

In this paper, we have introduced the P-MNER architecture, which has been specifically designed to tackle named entity recognition (MNER) tasks. Our proposed architecture leverages the power of prompt learning to process modal fusion features, thereby enabling the model to fully exploit the wealth of knowledge that pre-trained language models have to offer during training. We also proposed a fine-grained visual object feature extraction module (FRR) to address the issue of noise caused by irrelevant visual areas. This module aids in the MNER task by extracting only the relevant visual information, thus improving the accuracy of the model. To further address the issue of visual bias across modes, we proposed a new text correction fusion module. This module aligns the fusion features with text features to reduce visual bias and improve the model’s performance. Experimental results on benchmark datasets demonstrate that our P-MNER model outperforms state-of-the-art approaches. Our model’s superior performance is attributed to its ability to effectively utilize pre-trained language models and its innovative feature extraction and fusion modules. Overall, our proposed P-MNER architecture offers a promising solution for named entity recognition tasks, and we believe that our approach can be extended to other natural language processing tasks to improve their performance.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61976124)

## References

- Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2064.
- Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.
- Zijun Cui, Pavan Kapanipathi, Kartik Talamadupula, Tian Gao, and Qiang Ji. 2021. Type-augmented relation prediction in knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7151–7159.
- RCNN Faster. 2015. Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 9199(10.5555):2969239–2969250.
- Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Armand Joulin, Moustapha Cissé, David Grangier, Hervé Jégou, et al. 2017. Efficient softmax approximation for gpus. In *International conference on machine learning*, pages 1302–1310. PMLR.
- Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.
- Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.
- Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534.
- Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13860–13868.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022. Promptmer: Prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III*, pages 297–305. Springer.
- 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1038–1046.
- Zeguan Xiao, Jiarun Wu, Qingliang Chen, and Congjian Deng. 2021. Bert4gcn: Using bert intermediate layers to augment gcn for aspect-based sentiment classification. *arXiv preprint arXiv:2110.00171*.

- Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.
- Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In Proceedings of the AAAI conference on artificial intelligence, volume 32.
- Zehuan Zhao, Zhihao Yang, Ling Luo, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. 2017. Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC medical genomics, 10:75–83.
- Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, and Qing Li. 2020. Object-aware multimodal named entity recognition in social media posts with adversarial learning. IEEE Transactions on Multimedia, 23:2520–2532.
- Jingwei Zhuo, Yong Cao, Jun Zhu, Bo Zhang, and Zaiqing Nie. 2016. Segment-level sequence modeling using gated recursive semi-markov conditional random fields. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1413–1423.

JCL 2023