

MMAF: Masked Multi-modal Attention Fusion to Reduce Bias of Visual Features for Named Entity Recognition

Jinhui Pang^{1*}, Xinyun Yang¹, Xiaoyao Qiu², Zixuan Wang¹, Taisheng Huang¹

¹Beijing Institute of Technology, Beijing 100081, China

²Imperial College London, London SW7 2AZ, United Kingdom

Keywords: Multi-modal named entity recognition; Multi-head attention; Reduce bias of visual features; Image caption; Attention mask

Citation: Pang J.H., Yang X.Y., Qiu X.Y., et al.: MMAF: Masked Multi-modal Attention Fusion to Reduce Bias of Visual Features for Named Entity Recognition. *Data Intelligence* 6(4), 1114-1133 (2024). doi: <https://doi.org/10.3724/2096-7004.di.2024.0049>
Submitted: February 11, 2024; Revised: April 11, 2024; Accepted: August 20, 2024

ABSTRACT

Multi-modal Named Entity Recognition (MNER), which is vision-language task, utilizes images as auxiliary to detect and classify named entities from input sentence. Recent studies find visual information is helpful for Named Entity Recognition (NER), while the difference between those two modalities is not carefully considered. Therefore, these approaches utilizing different pre-trained models do not reduce the gap between textual and visual features, which give the same weight of different modalities usually predict wrong because of the noise of visual information. To reduce these bias, we propose a Masked Multi-modal Attention Fusion approach for MNER, named MMAF. Firstly, we utilize Image Caption to generate textual representation of image, which is combined with original sentence. Then, to get textual and visual features, we map the multi-modal inputs into a shared space and stack Multi-modal Attention Fusion layer that performs fully interaction between two modalities. We add Multi-modal Attention Mask to highlight the importance of certain words in sentences, enhancing the performance of entity detection. Finally, we achieve Multi-modal Attention based representation for each word and perform entity labeling via CRF decoder. Experiments show our method outperforms state-of-the-art models by 0.23% and 0.84% on Twitter 2015 and 2017 MNER datasets respectively, demonstrating its effectiveness.

* Corresponding author: Jinhui Pang (Email: pangjinhui@bit.edu.cn; ORCID: 0000-0002-7946-7089).

1. INTRODUCTION

Nowadays, Multi-modal Named Entity Recognition (MNER) has attracted extensive attention in natural language processing, is a sub-task of information extraction in NLP, which aims to recognize some entities with specific meanings from a large amount of text information. It mainly includes three major categories (entity, time, and number) and seven sub-categories (person, place, organization, time, date currency and percentage). On the basis of traditional text-based NER, MNER alleviates ambiguity in natural language through utilizing auxiliary image information, which is firstly introduced for noisy user-generated data such as tweets [1]. As social media posts become multi-modal, it becomes more and more challenging to understand short text messages on social media platforms, where text-only messages will lead to ambiguity in comprehension. For example, in Figure 1(a), without the picture as a prompt, it's hard to judge Jordan is whether a shoe brand or a person. And in Figure 1(b), we recognize Herman as a person with text-only message, where we can know that Herman is actually name of a snake with the image information which should be recognized as MISC entity. Therefore, we can refer that image information is important, which could helps us understand semantics better.



Kobe [PER] is playing basketball
wearing Jordan [MISC].

(a)



Herman [MISC], is enjoying the sun
in the afternoon!

(b)

Figure 1. Two examples for multi-modal named entity recognition with better semantic understanding aided by images^①.

^① These two examples are from Twitter 2015.

Apparently, how to exploit visual information and fully fuse with textual information is one of major problems of MNER task. Early works encode the whole image as a global feature [1, 2, 3, 4], then researchers use the processed images, such as segmenting the whole image into multiple regions [5, 6], converting the image into text [9] and vice versa. Although these models have some processes, they ignore either the fine-grained semantic correspondence or the gap between different modalities. The other problem is how to balance the relative importance of multiple modalities, researchers first utilize the attention mechanism [7, 42] to model the interaction between visual and textual representations implicitly [5, 6, 8], and then propose a Cross-Modal Alignment Module [9, 10]. However, these models do not adequately integrate information and balance the relative importance between the modalities.

To handle the above challenges, we propose a Masked Multi-modal Attention Fusion approach for MNER. We first convert the whole image into text representations and concatenate the sentence and the image caption. Next we input these text and region images into Multi-head Self Attention layer to extract the important features, and propose the Multi-modal Attention Mechanism, which can reduce the irrelevant information and the noise from images. We input the sentence and region images into one of the Multi-modal Attention layer. In the Multi-modal Attention layer, we propose a Multi-modal Attention Mask, which can enhance the correlation among the words, and ensure fully fusion between different modalities. After the attention layers, we can get the final textual and visual representations. Finally, we only exploit textual representations with a linear-chain CRF decoder to perform entity labelling. Compared with the previous works, our approach is able to fully exploit the fine-grained semantic correspondence between semantic units of different modalities and model the interaction between the aligned visual and textual representations with the Multi-modal Attention Mechanism for MNER. Overall, the major contributions of our work are listed as follows:

- We propose a Masked Multi-modal Attention based framework for MNER to model the interaction between textual and visual modalities, which adds a Multi-modal Attention Mask into Multi-modal Attention layer to balance the importance of different modalities and reduce the noise from the image.
- We propose the concatenation of the Image Caption and the original sentence, then input the sentence with the region images, where the image information could be fully exploited and the image caption could align the output distributions between textual and visual features.
- We conduct extensive experiments on two public MNER datasets, Twitter 2015 and Twitter 2017. Experimental results show that MMAF outperforms the current state-of-the-art models on both two datasets.

2. RELATED WORK

NER has attracted much attention as a crucial component of information extraction. It has been applied to lots of domains such as news [11], E-commerce [12], social media [13] and bio-medicine [14]. Nowadays, researchers focus on improving the accuracy of NER models through utilizing image information in tweets. And visual information could be aligned into same space as textual information by Image Caption. Therefore, we mainly overview MNER and Image Caption task in the following.

MNER. How to make better use of visual information is important in MNER. Yu segmented the whole image averagely into multiple regions and proposed a Multi-modal Interaction Module (MMI) based on Cross-Modal Attention. Zhang detected bounding boxes for each noun phrase from the input sentence using visual grounding toolkit and proposed a unified multi-modal graph fusion approach (UMGF), in which extended GNN, using the encoding of nodes [43] to model the interaction of different modalities. Sun proposed a Relation Propagation to select visual clues correspond more to the text. Wang converted visual information into textual information using VinVL [15] for object tags and image caption and Tesseract OCR [16] for OCR. To align the representations between the two modalities, Xu proposed a Cross-Modal Alignment Module, using cosine similarity between visual and textual features as contrastive loss. To optimize the learning of the text-image shared latent semantic space, Li proposed a de-bias contrastive learning based approach for MNER [17]. Currently NER task is still recognized as a sequence labeling problem in most literatures, which is only suitable for flat NER and unsuitable for nested NER where a token may be assigned several labels. To handle the problem, researches [18, 19] first formulated NER as a machine reading comprehension (MRC) task. And inspired by the this work, Jia [20] first applied the MRC framework to the MNER task, utilizing prior information to take advantage of MRC.

Different from above studies, we conduct the Multi-modal Attention layer to balance the weight of different modalities, where textual features are more important, to avoid encoding irrelevant information in visual features. We conduct Multi-modal Attention Mask added into Multi-modal Attention layer, which conducts the connection among words and full fusion between different modalities.

Image Caption. Image Caption has been a fundamental task in vision language understanding, where the model generated a corresponding textual description for the input image. BUTD [21] proposed a bottom-up and top-down attention mechanism to calculate the attention weights between objects and other salient image regions. Inspired by the success of pre-trained language model, Vision-Language Pre-training (VLP) model [22] utilize an Unified Encoder-Decoder model for Image Captioning and Visual Question Answering. OSCAR [23] represented each input image-text pair as a Word-Tag-Image triple to solve the problem of region overlap and lack of grounding. Vision-language pre-training (VLP) research has witnessed a rapid advancement in the past few years, where pre-trained models with increasingly larger scale have been developed to continuously push the state-of-the-art on various downstream tasks [24, 25, 26, 27], perform well on Image Caption task. Like Large Language Models(LLM), VLP took large scale datasets and too much time to train, which is not suitable for our MNER task.

In our work, we utilize ClipCap [28], which utilizes CLIP [29] as the encoder and GPT2 [30] as the decoder, to generate the textual description of the input sentence. And it achieves comparable results while it needs less time to train.

3. THE PROPOSED APPROACH

In this section, we introduce a Masked Multi-modal Attention Fusion model for MNER (MMAF), which is shown in Figure 2. We first introduce the task definition, then illustrate the architecture and describe its training procedure in detail.

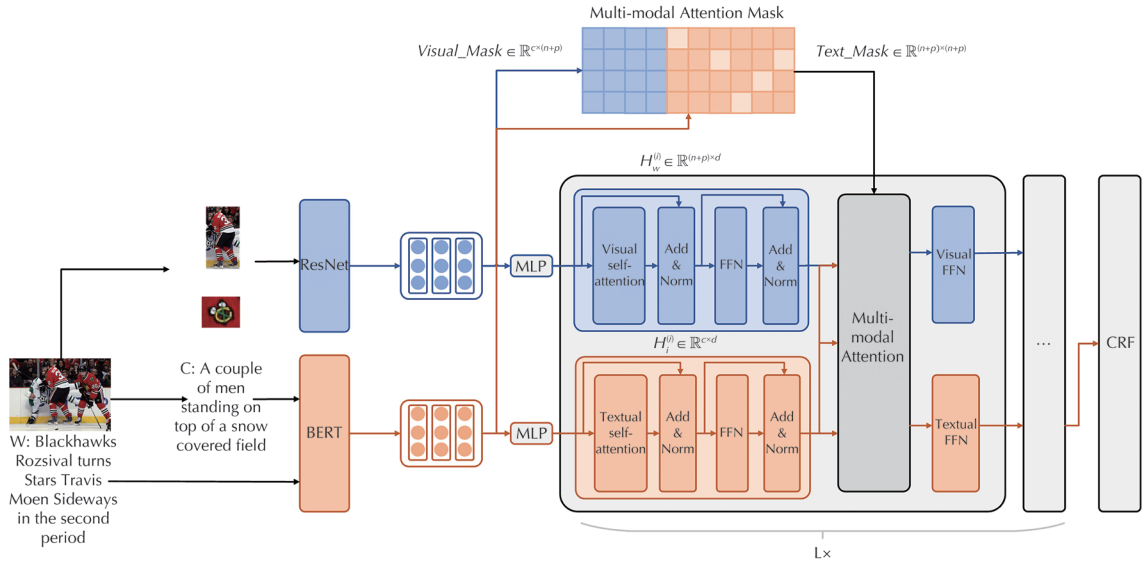


Figure 2. The architecture of MMAF. MMAF aligns image into region images and image captions. MMAF first takes image captions as visual contexts, concatenating with the input sentences and feeds them into BERT together while feeding the region images into ResNet. Next we utilize the output attentions from BERT to build the Multi-modal Attention Mask. Then we input the Multi-modal Attention Mask into Multi-modal Attention layer to fully fuse the textual and visual features more corresponding. Finally we input the textual features into CRF layer to decode.

Task Definition. We consider the MNER task as a sequence labeling problem. Given a sentence $W = \{w_1, \dots, w_n\}$ with n tokens and its corresponding Image I , the goal of MNER is to extract a set of entities from W , and classify each extracted entity into one of the pre-defined types. Let $W = \{w_1, \dots, w_n\}$ denote a sequence of input words, and $y = \{y_1, \dots, y_n\}$ be the corresponding label sequence, where $y_i \in Y$ and Y is the pre-defined label set with the BIOES tagging schema [31].

3.1 Data Preparation

We take the sentence and the whole image as original input. We will describe how to utilize the original image to generate the region images and the image caption. And then we will describe how to generate the textual and visual features.

Textual Features Construction. We utilize the input words and image caption as the final input sequence. The input sequence of our MMAF is designed as follows:

$$[\text{CLS}] \underbrace{w_1 \dots w_n}_W [\text{SEP}] \underbrace{c_1 \dots c_m}_C \quad (1)$$

where [CLS] stands for text classification, [SEP] stands for the separation between text and image caption features, $W = \{w_1, \dots, w_n\}$ denotes the sequence of original text features, and $C = \{c_1, \dots, c_m\}$ denotes the

sequence of Image Caption generated from the original whole image features. We utilize CILpCap [28] to generate the textual representation of the whole image. We generate the output, which is the textual representation of the original image. And then we concatenate the original sentence and the image caption sentence to get the final input. For example, in Figure 2, we can conduct the input sentence: “[CLS] Blackhawks Rozsival turns Stars Travis Moen Sideways in the second period [SEP] A couple of men standing on top of a snow covered field”. In the above sentence, Image Caption sentence adds to the original sentence, “A couple of men” focuses more attention on entity “person”, so that we would easier know that “Rozsival” and “Travis Moen” belong to entity “person”. Finally we generate the word token sequence by the BERT [32] tokenizer, which breaks an unknown word into multiple word-piece tokens.

Visual Features Construction. To capture the fine-grained semantic features in I , instead of employing the whole image as input, we need to derive additional visual objects from the text. Similar to UMGF [6], we employ the Stanford parser to identify all noun phrases in the input sentences, and then apply visual grounding toolkit [33] to detecting bounding boxes (visual objects) for each noun phrase. Instead of detecting all related objects from the whole image, we introduce four general words of pre-defined entity types of NER (i.e., person, location, organization and others) to only encourage corresponding objects. Since we dividing the whole image into four region images, we input them into ResNet [34] to extract visual features. Object detection exploits the fine-grained semantic correspondences and we detect the objects related to words, which reduce the irrelevant information from the image.

3.2 Masked Multi-modal Attention Fusion

As shown in Figure 2, we extract textual and visual features utilizing BERT and ResNet. Before inputting these features into our Multi-modal Attention Fusion layer, we introduce two multi-layer perceptrons (MLP) with ReLU activation function to project two different modalities onto the same space. For textual features, we form a textual MLP which is a fully connected layer to form dimension d . For visual features, we form a visual MLP, which is conducted by two fully connected layers and a ReLU activation function, to form dimension d . After the MLP layer, we define the initial state $H_w^{(0)} \in \mathbb{R}^{(n+p) \times d}$ as the textual features matrix, where n represents the length of input sentence tokens, p represents the length of image caption tokens and d represents the dimension of each word features. Similarly, we define the initial state $H_i^{(0)} \in \mathbb{R}^{c \times d}$ as the visual features matrix, where c represents the number of region images and d represents the dimension of each image features.

After the MLP layers, we stack L Multi-modal attention layers to encode our textual and visual features. At each fusion layer, we sequentially conduct Multi-head Attention and Multi-modal Attention fusion to fully update all features. In this way, the final features encode both the context within the same modality and the cross-modal semantic information simultaneously. In particular, we also conduct a Multi-modal Attention Mask, which is used for confirming the connection among words while fully fusing between different modalities.

Specifically, in the l -th fusion layer, both updates of textual features states $H_w^{(l)}$ and visual features states $H_i^{(l)}$ mainly involve the following steps:

Multi-head Attention. Similar to UMGF [6], we employ Multi-head Self Attention mechanism, which is also the encoder part of transformer, to generate the representations of each modalities. Formally, the textual representations $C_w^{(l)} \in \mathbb{R}^{(n+p) \times d}$ of all textual features are calculated as follows:

$$C_w^{(l)} = \text{Multihead}(H_w^{(l)}, H_w^{(l)}, H_w^{(l)}) \quad (2)$$

where Multi-head Attention function which needs a query matrix Q , a key matrix K , and a value matrix V as inputs. We utilize the same feature matrix $H_w^{(l-1)}$ to generate Q , K and V . Then we generate $C_w^{(l)}$ through the Multi-head Attention layer. Similarly, the visual representations $C_i^{(l)} \in \mathbb{R}^{c \times d}$ are calculated like eq.2. Next, we use textual features as an example, stacking FFN and norm layer as follows:

$$\begin{aligned} \widetilde{R}_w^{(l)} &= \text{LN}(H_w^{(l-1)} + C_w^{(l)}) \\ R_w^{(l)} &= \text{LN}(\widetilde{R}_w^{(l)} + \text{FFN}(\widetilde{R}_w^{(l)})) \end{aligned} \quad (3)$$

where FFN is the feed-forward network, LN is the layer normalization, visual features $R_i^{(l)}$ are generated like above.

Multi-modal Attention Fusion. We employ Multi-modal Attention mechanism [35] to consider the relative importance of multiple modalities. In our Multi-modal Attention layer, we utilize both textual features $R_w^{(l)}$ and visual features $R_i^{(l)}$ as input to generate new representation. Formally, the representation of query matrix $Q \in \mathbb{R}^{(n+p+c) \times d}$, key matrix $K \in \mathbb{R}^{(n+p) \times d}$ and value matrix $V \in \mathbb{R}^{(n+p) \times d}$ is calculated as follows:

$$\begin{aligned} Q &= [R_w^{(l)}, R_i^{(l)}] W_q \\ K &= R_w^{(l)} W_k \\ V &= R_w^{(l)} W_v \end{aligned} \quad (4)$$

where $[\cdot]$ is concatenation of $R_w^{(l)}$ and $R_i^{(l)}$. Formally, the output representation of our Multi-modal Attention is calculated as follows:

$$Z^{(l)} = \text{MultiHead}(Q, K, V, \text{Attention_Mask}) \quad (5)$$

where $Z^{(l)} \in \mathbb{R}^{(n+p+c) \times d}$ is the output of Multi-modal Attention layer, next we divide $Z^{(l)}$ into textual features $Z_w^{(l)}$ and visual features $Z_i^{(l)}$, and we will introduce the *Attention_Mask* below. Then we stack FFN layer like eq.3 to generate $H_w^{(l)}$ and $H_i^{(l)}$ as the input of next layer.

Multi-modal Attention Mask. The Multi-modal Attention Mask is formally encapsulated within the framework of the Multi-modal Attention mechanism as eq.5. In this equation, the *Attention_Mask* is our specific Multi-modal Attention Mask. Our methodology involves feeding both textual and visual features into the Multi-modal Attention layer. The Multi-modal Attention masking strategy for each objective is to control word-word and vision-text interaction. Different words have different contextual associations with each other, so we utilize the output attentions from BERT, to generate the *Text_Mask* $\in \mathbb{R}^{(n+p) \times (n+p)}$ among textual features, where represents the connection among $(n + p)$ words to highlight key words. And we conduct *Visual_Mask* $\in \mathbb{R}^{c \times (n+p)}$, where represents the the connection between $(n + p)$ words and c Region images. Each element value of the *Visual_Mask* is 1 because we need words and images to see each other, we cannot ignore any input regions. Therefore, We conduct full connection to ensure full fusion of textual and visual features in Multi-modal Attention layer, where also balances the weight of different modalities.

Finally, we conduct our *Attention_Mask* = [*Text_Mask*; *Visual_Mask*], which is a member of the set of real numbers and has the dimensions $\mathbb{R}^{(n+p+c) \times (n+p)}$. This *Attention_Mask* is then fed into our Multi-modal Attention layer to ensure fully fusion between textual and visual features and confirm correlation within textual features.

3.3 CRF Decoding

Since we conduct our Multi-modal Attention layer to fully fuse textual and visual features, we utilize the Conditional Random Fields (CRF) [36] layer to perform sequence labelling task. We only extract textual features $H_w^{(L)}$ as the final hidden representations and feed them into a linear-chain CRF layer to predict the the probability of a sequence of predictions Y through the original text W and its associated image I . We define it as follows:

$$p(y | W, I) = \frac{\prod_{i=1}^n \mathcal{S}_i(y_{i-1}, y_i, W, I)}{\sum_{y' \in Y} \prod_{i=1}^n \mathcal{S}_i(y'_{i-1}, y'_i, W, I)} \quad (6)$$

where $\mathcal{S}_i(y_{i-1}, y_i, W, I)$ and $\mathcal{S}_i(y'_{i-1}, y'_i, W, I)$ are potential score functions.

To train our module, we use the maximum log-likelihood loss as our loss function, which is defined as follows:

$$L(p(y | W, I)) = \sum_j \log p(y | W, I) \quad (7)$$

We use maximum log-likelihood estimation to learn parameters to maximize the maximum log-likelihood loss $L(p(y|W,I))$. In the decoding phrase, we predict the output sequence that obtains the maximum score, which is defined as follows:

$$\hat{y} = \operatorname{argmax}_{y' \in Y} p(y | W, I) \quad (8)$$

4. EXPERIMENTS AND ANALYSIS

We conduct experiments on two public datasets (Twitter 2015 and Twitter 2017), comparing our Multi-modal Attention Fusion(MMAF) with other Uni-modal and Multi-modal NER approaches (Table 1).

Table 1. Performance comparison on significant test $p < 0.05$ of some competitive uni-modal and multi-modal NER approaches.

Approaches	Twitter-2015							Twitter-2017						
	Single Type (F_1)				Overall			Single Type (F_1)				Overall		
	PER	LOC	ORG	MISC	Pre.	Rec.	F_1	PER	LOC	ORG	MISC	Pre.	Rec.	F_1
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
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
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
RpBERT	-	-	-	-	-	-	74.40	-	-	-	-	-	-	-
MAF	84.67	81.18	63.35	41.82	71.86	75.10	73.42	91.51	85.80	85.10	68.79	86.13	86.38	86.25
MRC-MNER	85.71	81.97	61.12	40.20	78.10	71.45	74.63	92.64	86.47	83.16	72.66	88.78	85.00	86.85
RDS	-	-	-	-	71.96	75.00	73.44	-	-	-	-	86.25	86.38	86.32
GEI	-	-	-	-	73.39	75.51	74.43	-	-	-	-	86.16	87.49	86.82
Ours	86.63	82.27	64.17	42.83	74.15	76.06	75.08	93.25	86.46	84.76	74.75	88.25	87.20	87.72

Note: We refer the results of uni-modal approaches from UMGF [6], and the results of multi-modal approaches are retrieved from the corresponding papers.

4.1 Experimental Settings

Datasets. We evaluate our framework on two public MNER datasets, Twitter 2015 [3] and Twitter 2017 [2]. Each dataset is conducted by sentence-image pairs which contains four entity types: person (PER), organization (ORG), location (LOC) and others (MISC). Each dataset is divided into a training set, a validation set and a test set with the same type distribution and the statistics summary are listed on Table 2.

Table 2. The Statistics Summary of Two MNER Datasets.

Type	Twitter 2015			Twitter 2017		
	Train	Dev	Test	Train	Dev	Test
PER.	2217	552	1816	2943	626	621
LOC.	2091	522	1697	731	173	178
ORG.	928	247	839	1674	375	395
MISC.	940	225	726	701	150	157
Total	6126	1546	5078	6049	1324	1351
Tweets	4000	1000	3257	3373	723	723

Evaluation Metrics. Following the previous work [5, 6], we utilize precision (Pre.), Recall (Rec.), and F1-score (F1) to evaluate the performance of Multi-modal Named Entity Recognition for overall entity types, and only use F1-score to evaluate each type.

Implementation Details. For our work, we refer to some UMGF [6] settings. We set the maximum length of the input sentence 128. The word embedding W is initialized with the BERT model [32] with the dimension of 768 and fine-tuned during training. The visual embedding is initialized with the ResNet152 [34] with the dimension of 2048 and fine-tuned during training. These features are transformed to the dimension d which is 512 by the MLP layer. The number of heads in multi-head attention layer is set to 8. Different to UMGF, we set the batch size to 8. The learning rate and the dropout rate are set to $1e-4$ and 0.5, which can achieve the best performance on the test set of both datasets after conducting a grid search over the interval $[1e-5, 1e-3]$ and $[0.1, 0.6]$. To ensure textual and visual features can be fully fused while taking less time, the number of our Multi-modal Attention Fusion layer L is set to 3. The entire model is trained on one NVIDIA GeForce RTX 3090 with pytorch 1.7.1.

4.2 Baselines

In order to compare fully, we mainly compare two groups of baseline systems with our approach.

The first group contains some representative text-based NER models which formalize NER as a sequence labeling task: (1) **CNN-BiLSTM-CRF** [37], utilizes the classic BiLSTM-CRF network with both word embedding and character representation by CNN. (2) **HBiLSTM-CRF** [38], which is also an improvement of BiLSTM-CRF, gives the word embedding to a bidirectional LSTM and concatenate these two vectors yields the representation of the word in its context. (3) **BERT** [32], is the most competitive baseline for NER which utilize the pre-trained multi-layer transformer as encoder and stack a softmax layer for entity prediction. (4) **BERT-CRF**, which is an improvement of BERT for NER task, utilizes CRF to replace the softmax layer for entity prediction.

The second group contains some competitive Multi-modal NER approaches: (5) **UMT** [5], stacked a Multi-modal Interaction Module based on transformer and an Auxiliary Entity Span Detection Module to alleviate the visual bias. (6) **UMGF** [6], encoded word and visual object into nodes and conducts a unified multi-modal graph to fuse these textual and visual nodes. Some of the setting can be found in Implementation Details. (7) **RpBERT** [8], proposed a text-image relation propagation based multi-modal BERT model to reduce the interference from irrelevant images and leverage more visual information for relevant text-image pairs. (8) **MAF** [10], proposed a Cross-Modal Matching and Alignment Module to alleviate the mismatches and the semantic gap between text and image pairs. (9) **MRC-MNER** [20], which utilized entity types as prior information to query the region images and use query and sentence as text-modality, considered MNER as an MRC task to label both flat and nested entity. (10) **RDS** [39], proposed a general data splitting strategy and implement this strategy by reinforcement learning so that these two sets can achieve better performance under the information extraction models of the corresponding modalities. (11) **GEI** [40], proposed an end-to-end heterogeneous Graph-based Entity-level Interacting model to rid non-entity tokens of the visual noise.

4.3 Experimental Results

To demonstrate the effectiveness and generalization of our approach, we conduct extensive single-domain and cross-domain experiments.

Single-domain Experiment. Table 1 shows the results of our MMAF and other competitive baselines in single-domain. From this table, we can see that:

1) From the uni-modal approaches, we can consider that **BERT-CRF** outperforms **HBiLSTM-CRF** by 2.64% and 3.07% on overall F1-score. Apparently, pre-trained approaches perform better than the conventional neural networks. This shows that pre-trained approaches will achieve better results on NER tasks, which also explain why Multi-modal NER approaches utilize pre-trained models.

2) From the uni-modal approaches, we can consider that **BERT-CRF** outperforms **BERT** by 0.49% on overall F1-score of Twitter 2015 and Twitter 2017. Apparently, **BERT-CRF** with CRF decoder performs better than **BERT** with softmax, which shows that CRF is effective on NER task.

3) From the uni-modal and multi-modal approaches, we can consider that **UMT** outperforms **BERT-CRF** by 1.6% and 1.87% on overall F1-score, **UMGF** outperforms **BERT-CRF** by 3.04% and 2.07%, **MAF** outperforms **BERT-CRF** by 1.61% and 2.81% and **MRC-MNER** outperforms **BERT-CRF** by 2.82% and 3.41%. All of these multi-modal approaches utilize BERT-ResNet-CRF architecture and conduct a fusion layer of different modalities. Apparently, we can see that while utilizing the same pre-trained model architecture, the multi-modal approaches perform better than the uni-modal approaches. This demonstrates that adding the visual environment will facilitate NER task in some scenarios.

4) From the multi-modal approaches, we can consider that our **MMAF** outperforms **UMT** by 1.67% and 2.38% on overall F1-score, outperforms **UMGF** by 0.23% and 2.18%, outperforms **MAF** by 1.68% and

1.44%, outperforms **RpBERT** by 0.68%, outperforms **MRC-MNER** by 0.47% and 0.84%, outperforms **RDS** by 1.64% and 1.40% and outperforms **GEI** by 0.65% and 0.90%. While these approaches utilize BERT as the pre-trained model, our approach also uses BERT in order to ensure fairness and prove our approach performs better than others. So from the result we can consider that our approach performs better than others which demonstrates the advancement of our methods.

5) The recent approach **UMGF**, which utilizes visual grounding to detect objects related to entity type, constructing a graph connecting textual and visual features. Our **MMAF** can be seen an improvement of **UMGF**, using the same method to detect region images. However, we conduct the Multi-modal Attention Mask, which not only link the visual features and the textual features to fully fuse, but also highlight the connection among textual features, with a weight of [0, 1]. Our **MMAF** performs much better than **UMGF** on Twitter 2015 and Twitter 2017. Apparently, our approach is better than **UMGF**.

Cross-domain Experiment. Table 3 shows the performance comparison among **UMT**, **UMGF** and our **MMAF** in the Cross-domain Experiment [41] for generalization analysis. For the first part, Twitter 2015 ->Twitter 2017 denotes that we use the trained model on Twitter 2015 to test Twitter 2017. For the second part, Twitter 2017 ->Twitter 2015 denotes that we use the trained model on Twitter 2017 to test Twitter 2015. From the result we can see that our **MMAF** outperforms **UMT** by 3.6% and 1.56% on overall F1-score, **UMGF** by 1.52% on Twitter 2017 and yields competitive results on Twitter 2015. Therefore, our **MMAF** perform better than **UMGF** and **UMT**, which prove our approach is easier to transfer and has greater generalization capabilities.

Table 3. Performance comparison among UMT, UMGF and our MMAF in the Cross-domain Experiment for generalization analysis. We refer the results of UMT and UMGF from UMGF [6].

Approaches	Twitter 2015 ->Twitter 2017								Twitter 2017 ->Twitter 2015							
	Single Type (F ₁)				Overall				Single Type (F ₁)				Overall			
	PER	LOC	ORG	MISC	Pre.	Rec.	F ₁		PER	LOC	ORG	MISC	Pre.	Rec.	F ₁	
UMT	80.34	71.30	47.97	20.13	64.67	63.59	64.13		81.24	67.89	39.52	31.87	67.80	55.23	60.87	
UMGF	79.62	71.94	49.48	20.24	67.00	62.81	66.21		81.83	72.25	41.20	32.00	69.88	56.92	62.74	
MMAF (Ours)	82.58	73.07	52.27	25.62	69.89	65.69	67.73		82.59	68.69	41.39	32.85	67.82	57.84	62.43	

4.4 Analysis and Discussion

Ablation Study. To show the effectiveness of each component of our **MMAF**, we conduct ablation study and compare with the full **MMAF**:

- **MMAF w/o attm.**, a variant of our approach, which removes Multi-modal Attention Mask.
- **MMAF w/o mma.**, a variant of our approach, which removes Multi-modal Attention module.

- **MMAF w/o cap.**, a variant of our approach, which removes Image Caption as a part of textual input.
- **MMAF w/o reimg.**, a variant of our approach, which utilizes original image as visual input.

Table 4 shows the performance among the full **MMAF** and ablation study. First, compared with the full **MMAF**, the effect of **MMAF w/o attm.** is significantly reduced. Therefore, we can see that features of different modalities are fully fused in Multi-modal Attention layer by our Multi-modal Attention Mask, which also conducts the connection among textual features. Second, compared with the full **MMAF**, the effect of **MMAF w/o mma.** is significantly reduced. Therefore, we can see that textual and visual features are fully fused by our Multi-modal Attention layer, which also avoids encoding irrelevant information in images. Third, compared with the full **MMAF**, the effect of **MMAF w/o cap.** is significantly reduced. Therefore, we can see that Image Caption can be seen as a textual enhancement of image information, which maps images and text messages to the same space, explicitly. Finally, compared with the full **MMAF**, the effect of **MMAF w/o reimg.** is significantly reduced. Therefore, we can see that taking original image directly as input leads to ignoring much fine-grained semantic information, which could also result in inadequate fusion.

Table 4. Ablation study of our MMAF.

Approaches	Twitter 2015								Twitter 2017							
	Single Type (F_1)				Overall				Single Type (F_1)				Overall			
	PER	LOC	ORG	MISC	Pre.	Rec.	F_1		PER	LOC	ORG	MISC	Pre.	Rec.	F_1	
MMAF (Ours)	86.63	82.27	64.17	42.83	74.15	76.04	75.08	93.25	86.46	84.76	74.75	88.25	87.20	87.72		
MMAF w/o attm.	86.55	81.60	62.36	41.24	72.28	76.14	74.16	92.48	85.71	84.18	70.33	85.86	87.13	86.49		
MMAF w/o mma.	85.89	82.26	63.19	42.09	73.30	75.65	74.46	93.04	83.19	85.17	69.42	86.69	86.76	86.72		
MMAF w/o cap.	85.68	81.92	64.15	41.52	72.63	76.56	74.54	92.54	87.25	84.93	72.10	87.03	87.35	87.19		
MMAF w/o reimg.	86.19	82.08	62.47	42.45	72.06	76.70	74.31	91.80	84.83	83.91	72.23	85.28	87.05	86.16		

Attention Mask Robustness. We propose Multi-modal Attention Mask, which connects the textual and visual features related, and excludes interference from other irrelevant visual features. Therefore, we design Attention Mask Robustness Experiment to prove the noise resistance and robustness of our Multi-modal Attention Mask.

Table 5 shows the performance of noise resistance of our MMAF. We replace 20% of region images with random Gaussian noise while testing. We can see that without our Multi-modal Attention Mask, the result of testing with noisy images shows a significant drop in results compared to testing without noisy

images. However, our full **MMAF** performs stable, less affected by noisy images, which can show that our Multi-modal Attention Mask performs stable while more noise and bias is added to the images. We can show that our **MMAF** performs better in Robustness of noise perturbation due to our Multi-modal Attention Mask.

Table 5. Performance of noise resistance of our MMAF. We replace 20 percent of the images with random Gaussian noise.

Approaches		Twitter 2015								Twitter 2017							
		Single Type (F_1)				Overall				Single Type (F_1)				Overall			
		PER	LOC	ORG	MISC	Pre.	Rec.	F_1		PER	LOC	ORG	MISC	Pre.	Rec.	F_1	
no	MMAF (Ours)	86.63	82.27	64.17	42.83	74.15	76.04	75.08		93.25	86.46	84.76	74.75	88.25	87.20	87.72	
noise	MMAF w/o attm.	86.55	81.60	62.36	41.24	72.28	76.14	74.16		92.48	85.71	84.18	70.33	85.86	87.13	86.49	
with	MMAF (Ours)	86.62	82.26	64.05	42.83	74.13	76.01	75.06		92.96	85.95	85.50	73.24	87.45	87.64	87.55	
noise	MMAF w/o attm.	85.23	81.54	61.75	39.63	73.86	74.26	74.06		91.70	81.43	84.47	71.14	86.27	85.75	86.01	

Case Study. We conduct the Case Study with two special examples, in which our **MMAF** recognizes the entities correctly while **UMGF** fails, and parallel experiment to show the connection between images and sentences.

First, in Figure 3(a), we can see that although **UMGF** recognizes the second entity “Palestine” correctly, it gives the wrong prediction of the first entity. This is mainly because this approach pays too much attention on visual information about the place where the people are, while ignoring the word “Centre”. However, our MMAF utilize Multi-modal Attention Fusion, which balance the weight of textual and visual information. Our Multi-modal Attention Mask establishes connections between words, where the associations among different words vary, ensuring a complete fusion of different modalities. In this example, our approach balances bias and filters out noise from visual information, allowing our model to accurately follow the word “Centre” and correctly recognize it as the entity type “ORG”. Therefore, our **MMAF** performs better than **UMGF** because of our Multi-modal Attention Fusion.

Second, in Figure 3(b), we can see that although **UMGF** recognizes the first entity “Eric Lamaze” correctly, it gives the wrong prediction of the second entity. This is mainly because **UMGF** does not identify the meaning of “CoCo Bongo” and recognize it as the entity type “MISC”. However, in our approach, we utilize Image Caption, which concatenates with original sentence to generate textual input. So in this example, we can see that the word “CoCo Bongo” was connected to the words “dirt ground” in our Image Caption. Our approach aligns textual and visual features into the same space so that we

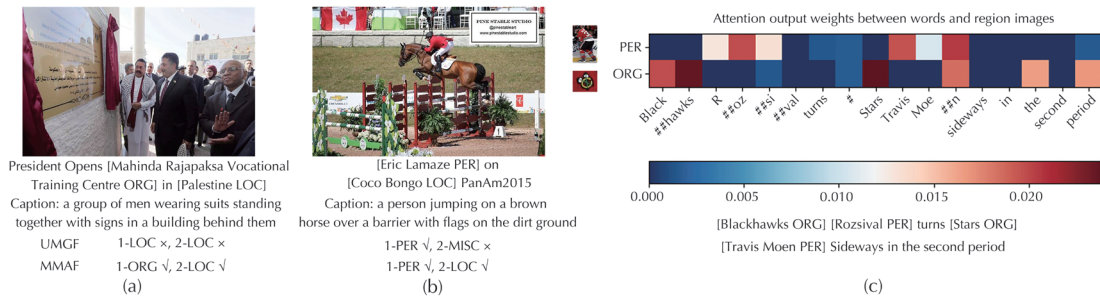


Figure 3. Examples of comparison of UMGF and our MMAF. Figures (a) and (b) demonstrate that our method outperforms UMGF, while Figure (c) illustrates a heat map of Multi-modal Attention between corresponding words and images.^②

recognize the word “CoCo Bongo” as the entity type “LOC” correctly. Therefore, our **MMAF** performs better than **UMGF** because of our Image Caption.

Third, in Figure 3(c), we did parallel experiment on Multi-modal Attention weights, where higher weights represent closer connection between corresponding words and images. And we can see that “Rozsival” and “Travis Moen” are given more “PER” weight and “Blackhawks” and “Stars” are given more “ORG” weight than others. Also, we can see that the word tokens “the” and “period” are given more “ORG” weight although they are type of non-identified entity, and the word token “n” belongs to “PER” type is also given much “ORG” weight. The “ORG” weight of these word tokens are still lower than the weights of real “ORG” word tokens. Although some other word tokens receive significant Multi-modal Attention weights, the word token corresponding to its entity type still captures more attention, indicating a closer connection.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a MMAF, a framework of Multi-modal Named Entity Recognition based on Masked Multi-modal Attention. We propose Multi-modal Attention Mask adding in the Multi-modal Attention layer to connect the visual and textual features correctly. We also utilize Image Caption, which turns images into its textual representation, to reduce the gap between different modalities.

In the future, we will try to distill more useful parts from the images to conduct our Multi-modal Attention Mask for the MNER task.

AUTHOR CONTRIBUTIONS

Xinyun Yang (3120221010.bit.edu.cn) was responsible for the system design, the experimental implementation of the approach, and the result analysis. Xiaoyao Qiu (qiuxiaoyao7@gmail.com)

^② Figures (a) and (b) are from our test set of Twitter 2015.

was responsible for the system design. Zixuan Wang (3220211081@bit.edu.cn), Taisheng Huang (h1820181076@gmail.com) were responsible for the result analysis. Jinhui Pang (pangjinhui@bit.edu.cn) were responsible for the system and research framework design. All authors contributed to the manuscript writing and approved the submitted version.

ACKNOWLEDGEMENTS

This work was supported by the Beijing Natural Science Foundation under Grant No.L247008.

REFERENCES

- [1] Seungwhan, M., Neves, L., Carvalho, V. Multimodal named entity recognition for short social media posts. arXiv preprint arXiv:1802.07862
- [2] Di, L.J.L., et al.: 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, pp. 1990–1999 (2018)
- [3] Zhang, Q., Fu, J., Liu, X., Huang, X. (2018). In Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, No. 1
- [4] Arshad, O., Gallo, I., Nawaz, S., Calefati, A. (2019). Aiding intra-text representations with visual context for multimodal named entity recognition. In: 2019 International conference on document analysis and recognition (ICDAR), pp. 337–342. IEEE
- [5] Yu, J., Jiang, J., Yang, L., Xia, R. (2020). Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: Association for Computational Linguistics
- [6] Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., Zhou, G. (2021). Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, No. 16, pp. 14347–14355
- [7] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems
- [8] Sun, L., Wang, J., Zhang, K., Su, Y., Weng, F. (2021). RpBERT: a text-image relation propagation-based BERT model for multimodal NER. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, No. 15, pp. 13860–13868
- [9] Wang, X., Gui, M., Jiang, Y., Jia, Z., Bach, N., Wang, T., ... Tu, K. (2021). ITA: Image-text alignments for multi-modal named entity recognition. arXiv preprint arXiv:2112.06482
- [10] Xu, B., Huang, S., Sha, C., Wang, H. (2022). MAF: a general matching and alignment framework for multimodal named entity recognition. In: Proceedings of the fifteenth ACM international conference on web search and data mining, pp. 1215–1223
- [11] Sang, E. F., De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050
- [12] Fetahu, B., Fang, A., Rokhlenko, O., Malmasi, S. (2021). Gazetteer enhanced named entity recognition for code-mixed web queries. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1677–1681
- [13] Derczynski, L., Nichols, E., Van Erp, M., Limsopatham, N. (2017). Results of the WNUT2017 shared task on novel and emerging entity recognition. In: Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 140–147

- [14] Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., ... Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database, 2016
- [15] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., ... Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5579–5588
- [16] Smith, R. (2007). An overview of the Tesseract OCR engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007) Vol. 2, pp. 629–633. IEEE
- [17] Zhang, X., Yuan, J., Li, L., Liu, J. (2023). Reducing the bias of visual objects in multimodal named entity recognition. In: Proceedings of the Sixteenth ACM international conference on web search and data mining, pp. 958–966
- [18] Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J. (2019). A unified MRC framework for named entity recognition. arXiv preprint arXiv:1910.11476
- [19] Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J. (2019). A unified MRC framework for named entity recognition. arXiv preprint arXiv:1910.11476
- [20] Jia, M., Shen, X., Shen, L., Pang, J., Liao, L., Song, Y., ... He, X. (2022). Query prior matters: A MRC framework for multimodal named entity recognition. In: Proceedings of the 30th ACM international conference on multimedia, pp. 3549–3558
- [21] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6077–6086
- [22] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 34, No. 07, pp. 13041–13049
- [23] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 2020, Proceedings, Part XXX 16, pp. 121–137. Springer International Publishing
- [24] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems, 34, pp. 9694–9705
- [25] Li, J., Li, D., Xiong, C., Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning, pp. 12888–12900. PMLR
- [26] Li, J., Li, D., Savarese, S., Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning, pp. 19730–19742. PMLR
- [27] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., ... Yang, H. (2022). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International conference on machine learning, pp. 23318–23340. PMLR
- [28] Mokady, R., Hertz, A., Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734
- [29] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763. PMLR

- [30] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9
- [31] Sang, E. F., Veenstra, J. (1999). Representing text chunks. arXiv preprint cs/9907006
- [32] Kenton, J. D. M. W. C., Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT, Vol. 1, p. 2
- [33] Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J. (2019). A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4683–4693
- [34] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778
- [35] Yao, S., Wan, X. (2020). Multimodal transformer for multimodal machine translation. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp. 4346–4350
- [36] Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Icml, Vol. 1, No. 2, p. 3
- [37] Ma, X. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. arXiv preprint arXiv:1603.01354
- [38] Lample, G. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv: 1603.01360
- [39] Xu, B., Huang, S., Du, M., Wang, H., Song, H., Sha, C., Xiao, Y. (2022, October). Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 1855–1864
- [40] Zhao, G., Dong, G., Shi, Y., Yan, H., Xu, W., Li, S. (2022). Entity-level interaction via heterogeneous graph for multimodal named entity recognition. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 6345–6350
- [41] Sekeroglu, B., Ever, Y. K., Dimililer, K., Al-Turjman, F. (2022). Comparative evaluation and comprehensive analysis of machine learning models for regression problems. Data Intelligence, 4(3), pp. 620–652
- [42] Zhang, Y., Chen, Y., Yu, S., Gu, X., Song, M., Peng, Y., ... Liu, Q. (2022). Bi-GRU relation extraction model based on keywords attention. Data Intelligence, 4(3), pp. 552–572
- [43] Chen, Z., Wu, Y., Feng, Y., Zhao, D. (2022). Integrating manifold knowledge for global entity linking with heterogeneous graphs. Data Intelligence, 4(1), pp. 20–40

AUTHOR BIOGRAPHY



Jinhui Pang received a bachelor's degree in Hunan University in 1991, and a Ph.D in Beijing Institute of Technology in 2008. She is director of Beijing Operations Research Society, Director of Game Theory Branch of Chinese Operations Research Society, and member of CCF. She has presided over provincial and ministerial projects such as the top project of the National Natural Science Foundation of China, the project of the Beijing Social Science Foundation, the project of the Humanities and Social Sciences Foundation of the Ministry of Education, etc. Her main academic research interests are game theory, complex decision-making and optimization methods, data mining and knowledge management, and statistical learning and applications.



Xinyun Yang received a bachelor's degree in computer science and technology from Beijing Institute of Technology in 2022, and studied for a master's degree in the same institution on September, 2022. His research interests include Natural Language Processing, Vision Language pre-training, Multi-modal Named Entity Recognition, etc.



Xiaoyao Qiu received his bachelor's degree in Artificial Intelligence from Beijing Institute of Technology in 2023. He attended to Imperial College Machine Learning Lab for Master of Research in Machine Learning from September 2023. His research focus on Generative Machine Learning, Score-based generative models and Stein's Method.



Zixuan Wang received a bachelor's degree in software engineering from Beijing University of Technology in 2021, and studied for a master's degree in Beijing Institute of Technology on September 2024, 2021. His research interests include GNN, Domain Adaptation, etc.



Taisheng Huang received the B.S. degree in Computer science and technology of Beijing Institute of Technology, Beijing, China in 2022 and during 2022.9-2024.7 get the Master's project in the same Institution. His research interests include AIGC, Graph and so on.