International Journal on Semantic Web and Information Systems
Volume 21 • Issue 1 • January-December 2025

Semantic Web-Enabled Multimodal Entity Recognition Using Collaborative Cross-Attention Mechanism

Dongxiu Wang
https://orcid.org/0009-0002-8171-1463
School of Economics and Management, Guangxi University of Science and Technology, Liuzhou, China

Yulan Wen
https://orcid.org/0009-0005-8392-6742
Wuzhou University, Wuzhou, China

Zhengxiang Qiu
Guangxi Boda Software Co., Ltd., Liuzhou, China

ABSTRACT

Named entity recognition (NER) is crucial in tasks such as information extraction, question and answer systems, and opinion analysis, but the existing methods still have deficiencies in cross-modal feature alignment and fusion, and it is difficult to make full use of visual information and structured knowledge to improve the recognition accuracy. To this end, this paper proposes CoAtt-NER, a multimodal NER model supporting semantic web, which combines textual representation generated by ALBERT with knowledge graph embedding to enrich the semantic information of entities; and adopts CLIP-ViT for better visual feature extraction. In addition, Co-Attention is proposed to establish two-way interaction between text and visual modalities to achieve dynamic modelling and deep fusion of information. Experiments on Twitter-2015 and Twitter-2017 datasets show that the F1 scores of CoAtt-NER reach 76.25% and 87.31%, respectively, which achieve significant improvement compared with existing methods, verifying the effectiveness of this study in multimodal entity recognition tasks.

KEYWORDS

Multimodal Named Entity Recognition, Collaborative Attention, Multimodal Fusion, Deep Interaction, Feature Extraction, Semantic Web

INTRODUCTION

With the continuous advancement of semantic web technologies and information processing methods, named entity recognition (NER) has evolved into a fundamental task within natural language processing, playing a crucial role in structuring and extracting meaningful knowledge from unstructured data. As a key component of the semantic web, NER enables the automatic identification and classification of entities—such as people, locations, and organizations—enhancing various applications, including information retrieval, intelligent question answering, and knowledge graph (KG) construction (Sang & De Meulder, 2003; Zhang et al., 2024). However, traditional NER approaches predominantly rely on text-only modalities, which limits their effectiveness in capturing rich, multimodal semantic information, making it difficult for unimodal NER models to achieve robust

DOI: 10.4018/IJSWIS.373310

and accurate entity recognition (Moon et al., 2018). The semantic web paradigm, which emphasizes structured knowledge representation and cross-modal integration, provides a promising framework for addressing these limitations by leveraging multimodal fusion techniques and linking extracted entities to structured knowledge bases. For example, the KG-enhanced NER approach improves recognition accuracy and contextual understanding by linking recognized entities to knowledge bases such as DBpedia and Wikidata. In addition, research based on joint visual-linguistic pre-training and multimodal graph neural networks has further explored the deep fusion of text, images, and structured knowledge, enabling NER models to perform better in multimodal scenarios, such as news, social media, and medical texts. Meanwhile, the ontology-driven NER approach ensures that the recognized entities conform to the conceptual hierarchy and reduces ambiguous parsing errors with the help of RDF/OWL semantic reasoning and SPARQL query.

To compensate for the limitations of single text modality, multimodal NER (MNER) has gradually emerged as a significant research direction in this field. In graphic content on social media, images often provide additional semantic information to text, especially in images involving entities or scenes, and visual modalities can significantly enhance the accuracy and robustness of entity recognition (Lai & Jie, 2023; C. Liu et al., 2024; F. Liu et al., 2019; X. Wang et al., 2021; L. Yang, 2024). For example, a masked multimodal attention fusion method has been proposed by Panga et al. (2024). This method converts images into text, integrates them with sentences, and uses multimodal attention to enhance text-visual interaction and entity recognition. However, masked multimodal attention fusion is deficient in feature extraction and cannot adequately capture contextual dependencies in text or high-level feature information in images. To address this problem, Zhang et al. (2024) suggested the contrastive language-image model, which enriches the ability of multimodal feature representation by enhancing useful information. Meanwhile, the approach introduces a hierarchical attention framework to facilitate the interaction of feature elements between different modalities. However, by relying only on the hierarchical attention framework, the method underperforms in the deep fusion of text and image features and fails to extract fine-grained interaction information between modalities.

Although previous studies have made progress in MNER, existing methods still encounter challenges such as insufficient unimodal feature extraction and inadequate fusion of text and image modalities. To address these issues, this paper introduces a semantic web-enabled MNER approach, CoAtt-NER, which integrates structured knowledge from the semantic web to enhance the accuracy and robustness of entity recognition. Unlike traditional MNER methods that rely solely on textual and visual features, CoAtt-NER incorporates KGs, a core component of the semantic web, to provide structured entity representations and contextual knowledge. Specifically, the model employs a lite bidirectional encoder representations from transformers (ALBERT) for text feature extraction and enhances it with KG embeddings, allowing for a deeper semantic understanding and disambiguation of entities. Meanwhile, for image feature extraction, the model utilizes CLIP-ViT, aligning visual representations with semantic knowledge from the KG to ensure that image-derived information contributes effectively to entity recognition. To achieve seamless multimodal fusion, a collaborative cross-attention mechanism dynamically models the interaction between textual and visual features while preserving their alignment with structured knowledge, enabling deeper cross-modal reasoning. By leveraging semantic web principles, CoAtt-NER effectively bridges the gap between structured knowledge and unstructured multimodal data, improving entity recognition performance, particularly in ambiguous or contextually limited scenarios. Experimental results demonstrate that the integration of semantic web technologies significantly enhances the effectiveness of MNER, outperforming existing state-of-the-art models.

The main innovations of this paper include the following. Focusing on the problem of insufficient feature extraction in traditional methods, this paper combines ALBERT and KG, which significantly improves the representation of text information through CLIP-ViT to enhance the accuracy and quality of image feature extraction. Focusing on the problem of insufficient interaction of inter-modal information, this paper also strengthens the synergistic cross-attention mechanism to enhance the

accuracy and quality of text and image features by focusing on the interaction information between text and images. depth interaction between modalities to ensure that the model can effectively capture the correlation between the two.

With the above improvements, the CoAtt-NER method suggested in this paper achieves significant performance improvement in the MNER task, especially when dealing with mixed graphical and textual information, demonstrating high accuracy and robustness. The innovation of the CoAtt-NER method in MNER not only improves the performance of specific tasks, but also provides powerful and intelligent support in the fields of fake news detection, cross-modal retrieval, and intelligent healthcare, which promotes the development of technology and solves practical problems.

Related Work

Unimodal NER

Unimodal NER is a significant research area in natural language processing that focuses on the identification and classification of specific types of entities, including the names of persons, places, organizations, and dates from text. Mo et al. (2023) suggested a cross-language NER approach, which transforms the traditional CrossNER task into a problem of recognizing relationships between token pairs, combines a multi-view comparison learning framework to reduce the gap between source and target languages, and significantly improves the cross-language NER effect especially when dealing with inter-language token pair relationship alignment. Chen et al. (2023) suggested a contextual learning approach that employs a limited number of examples and instructions to introduce NER capabilities into a pre-trained language model. The method enhances the efficacy of NER by dynamically generating entity recognizers and comparing pre-trained language model-generated entity extractors with standard extractors. Meng et al. (2022) introduced a NER model for news text based on bidirectional long short-term memory (Bi-LSTM) and conditional random field (CRF), which does not rely on manually labeled domain knowledge but captures contextual information through Bi-LSTM and optimizes the sequence labeling results by combining with CRF to improve the recognition accuracy. T. Yang et al. (2022) suggested a deep neural network-based NER method for medical text, which employs an attention mechanism to associate each word with the key information in the sentence, thus solving the ambiguity problem. The feature extraction efficiency is improved by multivariate convolutional decoding, which ultimately optimizes the recognition accuracy. Gao (2024) suggested an adaptive NER method for English, which combines the bidirectional encoder representations from transformers (BERT) pre-trained language model with BiLSTM-CRF, dynamically calculates features such as character and lexical properties, and enhances the recognition capability by splicing multiple features. Y. Ma et al. (2023) suggested to enhance the accuracy of NER by combining lexical information and local text features using convolutional neural networks and gated collaborative attention networks.

Sun et al. (2025) proposes a HREB-CRF framework that uses a hierarchical reduction of bias EMA for CRFs. The proposed method amplifies word boundaries and pools long text gradients by exponentially fixing the bias-weighted average of local and global hierarchical attention.

Compared with the MNER method suggested in this paper, unimodal NER has the advantages of simple implementation and lower computational resource requirements, but it is susceptible to the limitations of the modality itself, such as the lack of complementary and illustrative picture information, due to the reliance on only one source of information, which leads to inaccurate NER.

MNER

MNER is a multimodal, extended NER task that utilizes text, images, audio, or video to identify and classify specific categories of named entities, including the names of people, locations, organizations, and products. Unlike unimodal NER, which relies only on textual information, MNER is able to fuse data from different modalities, enabling the model to recognize entities more accurately in situations where text is ambiguous or contextually insufficient. For example, the CLMSI model

suggested in J. Ma et al. (2024) semantically aligns images and text by contrast learning to reduce the gap between them. Although the method can enhance semantic alignment through contrast learning, it only stops at surface alignment and fails to adequately model the deep interactions between images and texts. The TISGF method proposed in Cheng et al. (2024) utilized a scene graph to fuse the object and relationship features of text and images and dynamically adjusts the weights of visual information through text-image similarity assessment. Although it can realize information fusion, its similarity assessment is static and fails to dynamically adjust the weights of graphical and textual features according to the actual task. The MGICL approach proposed in Guo et al. (2023) enhances information alignment through cross-comparison learning by dividing text and images into different granularities for comparative learning, which in turn reduces the differences between them. The MGICL framework consists of sentence-level and lexical-element levels for text as well as whole- and object-levels for images. In addition, MGICL introduces a visual gating mechanism that dynamically selects the most relevant visual information to reduce noise interference. Despite the fact that the method enhances the accuracy of graphic and textual alignment through granularity segmentation, its contrast learning mechanism remains at the coarse-grained level and is unable to delve deeper into the fine-grained interaction information in images and texts. It may be unable to accurately convey the deep semantic relationships between images and texts, particularly when dealing with complex or detail-rich images. The MLNet approach in Zhai et al. (2023) performs feature fusion of text and images through the transformer architecture, combines a gating mechanism to filter text-related visual information, and performs entity recognition through CRFs. MLNet is effective in enhancing a semantic understanding of text and disambiguating ambiguities, but its approach still relies on a predefined gating mechanism, which may not be able to be dynamically implemented in different task scenarios. This static visual information filtering strategy makes the model unable to adaptively adjust the weights of different modal information according to the characteristics of the actual input data, which in turn affects the effectiveness of entity recognition. P. Liu et al. (2023) proposed a multimodal learning framework for dynamic alignment, which enhances the semantic alignment of text and images through cross-modal interactions and joint attention mechanisms. However, the approach relies on a static cross-modal alignment mechanism and fails to flexibly cope with changes between modalities, especially in scenarios where image information is decisive for NER, which may not fully utilize the rich information of images. Tourani et al. (2025) presents a novel real-time VSLAM framework that combines vision-based scene understanding with map reconstruction and comprehensible graph-based representation. The framework infers structural elements from detected building components and merges them into an optimizable 3D scene map. The solution enhances the semantic richness, comprehensibility, and localization accuracy of the reconstructed maps, improving the accuracy of environment-driven semantic entity detection.

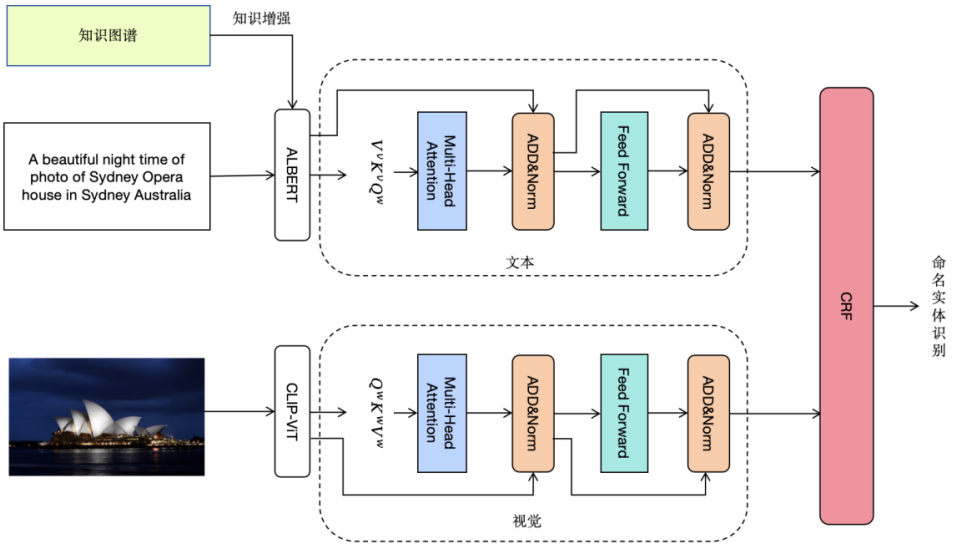
Unlike existing methods, the collaborative cross-attention mechanism suggested in this paper is able to dynamically adjust the weights between images and text to capture deeper interactive information between the two. Unlike the previous methods that only perform surface alignment or static fusion, our mechanism is able to optimize the contribution of image and text in real time according to the characteristics of each sample, ensuring that the fusion of image or text information always maximizes the accuracy of entity recognition in different modal environments. This deep interaction design enables our method to capture key information in images and text more accurately, thus effectively improving the performance of NER.

A Proposed MNER Method

The MNER approach studied in this paper extracts and fuses information from text and images for NER. The model structure of this paper is illustrated in Figure 1. The accuracy of NER is enhanced by using ALBERT with KG to enhance text feature representation, CLIP-ViT to extract image features, and collaborative cross-attention mechanism to deeply fuse text and image features. Given a sentence s and its corresponding image v, the objective of MNER is to recognize the entity E from the sentence s and

classify it into predefined types. The MNER model first extracts textual features $H = [h_1, h_2, \ldots, h_L]$ and visual features $E_T = \{E_{t_0}, E_{t_1}, \ldots, E_{t_m}\}$. Subsequently, these features are combined at the cross-modal fusion module to generate a cross-modal representation. Finally, the decoder assigns a corresponding label to each token based on the predefined set of entity labels Y of the MNER dataset.

Figure 1. The structure of model



Text Encoder

The ALBERT model (Lan et al., 2019), which is derived from BERT, is implemented to enhance the efficiency of training and decrease the number of parameters by utilizing matrix decomposition and parameter sharing. The innovations of ALBERT include word embedding and hidden layer unbinding, which reduces the number of parameters by mapping the word embedding dimension HEH_EHE to the hidden layer dimension HHH through matrix transformation. Another innovation is the sharing of the transformer layer parameters using the same parameters for each layer, avoiding the overhead of training each layer independently. Another innovation is using the SOP task, replacing the NSP task of BERT and focusing on the semantic relations of sentence order. The inputs to the model are first encoded by word embedding, positional embedding, and segmental embedding to obtain an initial representation, as shown in Eq. (1):

$$E = E_{word} + E_{pos} + E_{seg} \in \mathbb{R}^{L \times H_E} \tag{1}$$

To enhance the text representation, this paper integrates entity information into the KG in accordance with the ALBERT model. KG represents real-world semantic information in a structured way through nodes (entities) and edges (relationships), which can provide external knowledge support for text. By combining the entity embeddings in the KG with the word embeddings, positional embeddings, and segmental embeddings of the text, the missing information in the text can be effectively supplemented to enhance semantic understanding. Specifically, the entity embedding E_{KG} in the KG is extended and fused with the text embedding to obtain the joint representation $E_{combined} = E + \text{Expand}(E_{KG})$. This joint representation can provide richer semantic information for

subsequent model inputs, which helps the model to better utilize external knowledge when processing complex linguistic tasks. Then it is mapped to the hidden layer dimension by linear variation; this is followed by the input passing through the transformer encoder with shared parameters in the L_T layer and outputting the optimized text feature representation as H. Each transformer layer contains the multi-head attention mechanism and feed-forward neural network.

After L_T layers, the final text representation is obtained as $H = H^{(L_T)} = [h_1, h_2, \ldots, h_L] \in \mathbb{R}^{L \times H}$.

Visual Encoder

A visual transformer model based on contrast learning (Radford et al., 2021) is used in this paper. The difference with other visual transformer models is that CLIP-ViT is trained for multimodal prediction on a large number of image-text pairs, and its training data contains 400 million image-text pairs from the internet. This makes CLIP-ViT not only perform well on visual tasks, but also have strong cross-modal inference and zero-sample learning capabilities.

After CLIP-ViT processing, the input image is transformed into a feature embedding representation. For a given image embedding v, the specific form is shown in Eq. (2):

$$E_T = \left\{ E_{t_0}, E_{t_1}, \ldots, E_{t_m} \right\} = CLIP - ViT(v) \tag{2}$$

Where E_{t_0} is the global feature embedding of the image, $E_{t_1}, \ldots, E_{t_m}$ are the local feature vectors extracted from each patch of the image, and m represents the number of patches of the image.

After obtaining the corresponding embedding vectors of text and image through the text encoder and visual encoder, the collaborative cross-attention mechanism is used to realize the information interaction and sharing between cross-modal data.

The collaborative cross-focusing mechanism can effectively enhance the interactive information between text and images and improve the inter-modal fusion effect. The mechanism reduces information redundancy and improves the expression of important features by weighing and dynamically adjusting the focus of attention between modalities, thus improving the overall performance. Compared with other mechanisms, the collaborative cross-focusing mechanism is more flexible and able to adapt to different task requirements, and at the same time, it has higher robustness in dealing with mixed graphical and textual information.

Let the image and text features (embedding vectors) after encoding by the encoders are respectively, as shown in Eq. (3) and Eq. (4):

$$E_T = \left\{ E_{t_0}, E_{t_1}, \ldots, E_{t_m} \right\} \tag{3}$$

$$H = H^{(L_T)} = [h_1, h_2, \ldots, h_L] \tag{4}$$

The collaborative cross-attention mechanism takes E_T and E_I as inputs, and for the interaction from image to text, the visual feature E_{ij} is set as query and the text feature $H^{(L_T)}$ is set as key and value at the same time; and the cross-modal fusion of the text feature is represented as Eq. (5):

$$H^{(L_T)'} = CoAttention_{I \to T}\left(E_{ij}, H^{(L_T)} \right) \tag{5}$$

For the interaction from text to image, setting the text feature $H^{(L_T)}$ as query and the visual feature E_{ij} as both key and value, the cross-modal fused text feature is represented as Eq. (6):

$$E_I^{'} = CoAttention_{T \to I}\left(H^{(L_T)}, E_{ij} \right) \tag{6}$$

To avoid the over-concentration of attention on specific locations or features, which may occur with the traditional attention mechanism and lead to ignoring other important information, a smoothing attention score is introduced. Specifically, the attention score is smoothed by introducing a temperature parameter τ to control a more even distribution of attention. The temperature parameter τ controls the sharpness of the attention distribution: a higher τ makes the attention smoother, thus avoiding excessive attention to certain modalities or features and promoting a balanced fusion of information across modalities. This mechanism is especially suitable for modalities with different semantic spaces and structures, such as image and text, which avoids the model's over-reliance on a certain modality and enhances the stability and generalization ability of cross-modal learning. The smoothed attention score $\text{Score}_{I\to T}$ is calculated as Eq. (7):

$$Score_{I\to T} = softmax\left(\frac{\left(E_{ij} W_{I\to T}^{Q}\right)\left(H^{(L_7)} W_{I\to T}^{K}\right)^{T}}{\sqrt{d_k}\bullet \tau}\right) \tag{7}$$

Where τ represents a temperature parameter that controls the smoothness of the attention.

After the attention score is calculated, the features need to be weighed using the attention score to get the weighted text feature representation $E_i^{'}$, as shown in Eq. (8):

$$E_i^{'} = Score_{I\to T}\bullet \left(H^{(L_7)} W_{I\to T}^{V}\right) \tag{8}$$

W^{Q}, W^{K}, and W^{V} are the transformation matrices of query, key, and value in the image-to-text direction, respectively, and $\sqrt{d_k}$ represents the dimension of the key vector, which is employed for the scaling of the dot product result to stabilize the gradient.

All of the weighed text features $E_1^{'}, E_2^{'}, \ldots, E_L^{'}$ are composed into a new text feature representation as $H^{'} = \left[E_1^{'}, E_2^{'}, \ldots, E_L^{'}\right]$ as a cross-modal vector that is finally obtained by fusing the image and text data in the dataset, which is used as an input to the decoder.

Feature Decoding Layer

CRFs are used in this paper to decode cross-modal vectors of text. CRF is a statistical modeling method for predicting labels of sequence data. It is able to consider the dependencies between labels at the level of the whole sequence and find the optimal labeling configurations for the whole sequence instead of selecting the best labels for each position independently, and thus CRF has a significant advantage when dealing with the task of NER.

In the CRF layer, the label y_i at each sequence position i depends not only on the input features at that position but also on the labels at neighboring positions. This dependency is modeled by the transfer matrix A, where the matrix element $A_{k,l}$ represents the transfer score from label k to label l.

Given a cross-modal vector representation $H^{'} = \left[E_1^{'}, E_2^{'}, \ldots, E_L^{'}\right]$, where $E_i^{'}$ denotes the ith input feature in the sequence, the score function for the whole sequence is defined as Eq. (9):

$$Score\left(H', y\right) = \sum_{i=1}^{n}\left(W_{yi}\bullet E_i' + A_{y_{i-1}, y_i}\right) \tag{9}$$

Where W_{yi} is a vector of feature weights associated with label y_i and A_{y_{i-1}, y_i} is the transfer score from label y_{i-1} to label y_i.

In order to convert scores to probabilities, the scores on all possible label sequences $y^{'}$ need to be normalized, as shown in Eq. (10):

$$P(y|H') = \frac{\exp(Score(H', y))}{\sum_{y'}\exp(Score(H', y'))} \tag{10}$$

After that, the goal equation of the decoding process finds the most probable sequence label y^* given the input H', which is achieved by the Viterbi algorithm, which efficiently searches for the optimal path, as shown in Eq. (11):

$$y^* = \text{argmax}_y P(y|H') \tag{11}$$

The CRF layer can achieve high performance in the sequence annotation task in this paper by effectively considering the transfer probabilities between labels and providing accurate label predictions for text sequences represented across modal vectors through the aforementioned process.

EXPERIMENTS

Experimental Environment and Dataset

In this paper, the suggested model is trained with other baseline models in two widely used MNER datasets for model training, including the Twitter-2015 dataset (Q. Zhang et al., 2018) and the Twitter-2017 dataset (Lu et al., 2018). These two datasets cover social media texts with corresponding image information, making them standard testbeds for the MNER task. In data preprocessing, we removed samples with missing image files from the Twitter-2017 dataset in order to ensure the reliability of the experiments and the consistency of the data. Since MNER relies on the joint information of text and images, missing images may lead to incomplete features, thus affecting the learning effect of the model. Table 1 and Table 2 show an overview of the number of samples and entity types in each dataset division.

Table 1. Twitter-2015 dataset

Entity Type	Training set	Validation set	Test set
Person	2,217	552	1,816
Location	2,091	522	1,697
Organization	928	247	839
Other	940	225	726
Total	6,176	1,546	5,078
Number of Tweets	4,000	1,000	3,257

Table 2. Twitter-2017 dataset

Entity Type	Training set	Validation set	Test set
Person	2,294	626	621
Location	731	173	178
Organization	1,674	375	395
Other	701	150	157
Total	6,049	1,324	1,351
Number of Tweets	3,373	723	723

Experimental Setup

The main parameters of the experiments in this paper are configured as follows: Batch Size is 32, Optimizer is employed as Adam, Epochs are 40, Learning Rate is set to 0.001, and

Dropout is set to 0.1. These parameters are designed to ensure the stability and effectiveness of the model while striking a balance between the training efficiency and the risk of overfitting. The experiments were conducted in the hardware and software environments shown in Table 3, including NVIDIA GeForce RTX 3090 Ti (24GB) GPU, Intel Core i7-8750H @ 2.20GHz CPU, 1T hard disk space, and Anaconda3-Windows-x86_64 as the development environment. The programming language of the model is Python 3.10, and the development framework is selected as TensorFlow 1.14.0. For model performance evaluation, precision, recall, and F1 score are used as the main evaluation metrics to comprehensively measure the model's classification ability and prediction effect.

Precision measures the proportion of actual positive classes when the model predicts a positive class, reflecting whether the model is accurate in predicting positive classes; recall indicates the proportion of all samples that are actually positive classes that the model is able to correctly identify, reflecting the model's ability to capture positive samples; and F1 score is the reconciled average of precision and recall, which is used to comprehensively evaluate the model's performance in identifying the positive classes. F1 score is the reconciled average of precision rate and recall rate, which is used to comprehensively evaluate the overall performance of the model in identifying positive classes, especially in the case of data imbalance. By considering these three metrics together, the classification accuracy and robustness of the model in multimodal tasks can be reflected more comprehensively.

Model Performance Comparison

Experiments utilize the publicly available Twitter-2015 dataset, Twitter-2017 dataset, and the relevant MNER models are GDN-CMCF (Huang et al., 2024), PCEN (Geng et al., 2023), M3S (J. Wang et al., 2023), and TISGF (Cheng et al., 2024):

1. GDN-CMCF references a gated de-entanglement module, which separates text- and image-related features from support and assistance modalities and filters out irrelevant information.
2. PCEN employs unsupervised clustering to categorize training sample images into clusters, utilizing the trainable embeddings associated with each cluster as visual features. Additionally, textual features are configured with an inconsistency loss to assess the alignment of entity recognition outcomes for each sample with the distribution of pre-trained entity types.
3. M3S is a multimodal information-based approach to build scene graphs and gate control by constructing multi-granularity visual information so that the visual information is tightly integrated with the textual information.
4. TISGF generates two scene graphs (visual and textual) to capitalize on the shared characterization of objects and relationships in text and images. The two scene graphs are encoded separately using particular encoder pairs.

The performance comparison of the CoAtt-NER model with mainstream models on different publicly available datasets is illustrated in Table 3 and Table 4.

Table 3. Comparison of experimental results of Twitter-2015

Model	P (%)	R (%)	F1 score (%)
GDN-CMCF	71.69	74.40	74.04
PCEN	75.21	74.90	75.04
M3S	74.93	75.43	75.02

continued on following page

Table 3. Continued

Model	P (%)	R (%)	F1 score (%)
TISGF	71.13	75.33	73.18
CoAtt-NER	76.12	77.38	76.25

Note. P = precision; R = recall.

Table 4. Comparison of experimental results of Twitter-2017

Model	P (%)	R (%)	F1 score (%)
GDN-CMCF	85.48	85.94	85.72
PCEN	87.10	85.41	86.53
M3S	86.64	85.22	86.10
TISGF	83.21	85.42	84.36
CoAtt-NER	87.15	86.71	87.31

Note. P = precision; R = recall.

Compared with the best performing comparison model PCEN, the CoAtt-NER model demonstrates substantial superiority in several aspects. On the Twitter-2015 and Twitter-2017 datasets, this paper's model outperforms PCEN in terms of F1 value, recall, and precision, with the F1 value improved by 1.21% on Twitter-2015 and 0.78% on Twitter-2017. In addition, the recall and precision of this paper's model are also higher than those of PCEN, indicating that it improves both comprehensiveness and accuracy in recognizing entities.

Analyzing the reason for this, PCEN has a certain dependence on the construction of graph structure, and the model needs to dynamically construct the graph relationship at each input and transfer the information through the graph neural network. The construction of the graph needs to define the nodes and edges reasonably, and if the graph structure is not properly designed, it may lead to inaccurate information transfer, thus affecting the performance. In contrast, the model in this paper can handle inter-modal relationships more flexibly through the cross-modal collaborative cross-attention mechanism. Instead of explicitly constructing graph structures, the attention mechanism automatically captures inter-modal correlations through weight learning, resulting in greater flexibility and simplified design. In addition, with the cross-modal collaborative cross-attention mechanism, it can focus on fine-grained interactions between modalities and can efficiently fuse text and image features, reduce redundant information, and improve the fusion quality of cross-modal data.

Compared with the GDN-CMCF model, which uses a simple BERT model to obtain text features, the CoAtt-NER model uses ALBERT in combination with KG to enrich the semantic information of entities and improve the performance of text feature extraction. Compared with the M3S model, the CoAtt-NER model proposed in this paper uses a cross-modal collaborative cross-attention mechanism to better fuse text and image features, while M3S uses a simpler splicing fusion. Compared with the TISGF model, the TISGF uses two independent scene graphs modeled separately and then fused, which may lead to insufficient information sharing and limited cross-modal interaction, while the CoAtt-NER model adopts a cross-modal collaborative cross-attention mechanism to ensure the full fusion of the cross-modal features, thus improving the multimodal entity recognition effect.

Text Encoder Performance Comparison

The proposed model using a combination of ALBERT and KG is experimentally compared with four other text encoders—Glove, BERT, Word2vec, and ALBERT—on the Twitter-2015 dataset and the Twitter-2017 dataset; and the results are depicted in Table 5.

Table 5. Performance comparison of text encoders

Text Encoder	Twitter-2015	Twitter-2017
	F1 score (%)	F1 score (%)
+Word2Vec	72.81	82.73
+Glove	73.12	83.20
+BERT	75.10	85.71
+ALBERT	75.19	86.22
+ALBERT+Knowledge	76.25	87.31

Note. BERT = bidirectional encoder representations from transformers; ALBERT = a lite bidirectional encoder representations from transformers.

The results of the experiments in Table 5 indicate that the text encoder, which integrates ALBERT and KG within the CoAtt-NER model, has superior performance on both the Twitter-2015 and Twitter-2017 datasets. Specifically, on the Twitter-2015 dataset, ALBERT's F1 value of 76.25% exceeds the performance of the other four text encoders, which are 72.81% for Word2Vec, 73.12% for Glove, 75.10% for BERT, and 75.19% for ALBERT. On the Twitter-2017 dataset, ALBERT has an F1 value of 87.31%, again ahead of Word2Vec (82.73%), Glove (83.20%), BERT (85.71%), and ALBERT (86.22%).

The superior performance of the knowledge-enhanced ALBERT model on both datasets demonstrates that its lower number of parameters, efficient training method, and KG enhancement can provide a more accurate representation in text feature extraction, thus improving the overall performance of multimodal entity recognition. Compared with traditional Word2Vec and Glove, ALBERT+ KG has a greater ability to capture contextual dependencies and semantic details of textual information, while ALBERT+ KG further improves the efficiency of the model by decreasing the number of covariates and maintaining the performance advantage over BERT. In contrast to the ALBERT model, the addition of KG makes the model in this paper greatly enhance the capacity of text feature extraction, and by combining it with KG, ALBERT is able to better understand the semantic relationships in the text and enhance the ability of modeling entities and relationships. Therefore, ALBERT+ KG is a key advantageous component in this paper's model, significantly enhancing the efficacy of multimodal entity recognition tasks.

Ablation Experiment

To validate the efficacy of the various parts of the CoAtt-NER model for named entity detection, this paper conducted ablation experiments for relevant validation, removing several key parts of the model to verify their usefulness:

- w/o image features: In the proposed model, remove the image features and use only text features as feature output.
- w/o text features: In the proposed model, remove the text features and use only image features as feature output.

- w/o cross-modal feature fusion module: In the proposed model, the cross-modal fusion module is removed, and text and image features are output using direct splicing.

Table 6 illustrates the outcomes of the ablation experiments.

Table 6. Results of ablation experiments

Model	Twitter-2015			Twitter-2017		
	P	R	F1 score	P	R	F1 score
w/o image features	75.21	76.36	74.34	86.41	85.92	86.39
w/o text features	75.13	76.22	74.28	86.21	85.77	86.27
w/o cross-modal feature fusion module	74.21	75.12	73.01	85.68	85.29	85.31
CoAtt-NER	76.12	77.38	76.25	87.15	86.71	87.31

Note. P = precision; R = recall.

The ablation experiment findings presented in Table 6 demonstrate that the accuracy of the CoAtt-NER model proposed in this paper has been enhanced by the inclusion of all of the parts. This demonstrates the utility of all components of the model presented in this paper. Specifically, the F1 value of the model decreases the most when the cross-modal feature fusion module is removed—by 3.24% (Twitter-2015) and 2.00% (Twitter-2017). In contrast, the decrease in F1 value is smaller when removing image features and text features. The F1 values decreased by 1.91% (Twitter-2015) and 0.92% (Twitter-2017) when the image features were removed, while the F1 values decreased by 1.97% (Twitter-2015) and 1.04% (Twitter-2017) when the text features were removed. These results indicate that the cross-modal feature fusion module contributes the most to model performance, and its removal leads to the most significant performance degradation due to its ability to effectively integrate information from different modalities (e.g., text and image) to provide a more comprehensive feature representation. Text and image, as complementary modalities, provide semantic information and visual cues, respectively, and the model's capacity to identify named entities is enhanced by the fusion. In addition, cross-modal feature fusion enhances the robustness of the model and helps to compensate for missing information or noise interference in a single modality, thus enhancing the model's generalization capabilities. By capturing the semantic associations between text and images, the fusion module can understand the relationship between the two more precisely and improve the classification accuracy. At the same time, the fusion module can reduce redundancy and noise, optimize the model's performance, and automatically adjust the weights between modalities to ensure that the most valuable information receives more attention. As a result, the cross-modal feature fusion module provides a more comprehensive and accurate feature representation of the model, which significantly improves performance.

Visualization Case Study

To further demonstrate the advantages of the models in this paper, this experiment randomly selected two samples from Twitter's test dataset for case studies. At the same time, the two better models PCEN, M3S among the MNER models are selected as comparison models. Table 7 shows the comparison results of this paper's model and the comparison model on the two case samples.

Table 7. Case study

	Case 1		Case 2	
				
	[TWICE] go unnoticed in [Times Square] during 'TT' cover performance		[Oklahoma] legend [Bob stoops] reportedly had interest in an NFL job before retiring	
M3S	1-PER	2-PER	1-PER	2-PER
PCEN	1-PER	2-LOC	1-PER-	2-PER
CoAtt-NER	1-PER	2-LOC	1-ORG	2-PER

First, in Case 1, the M3S incorporates visual information indiscriminately into the textual representation, thus incorrectly dropping the 'Times Square' and 'LOC' entity to 'PER' because the image contains 'PERSON' information. This is because M3S fails to distinguish the relevance of visual information in the multimodal fusion process, resulting in images with 'PERSON' information interfering with the correct recognition of text entities. In contrast, PCEN and CoAtt-NER correctly recognize 'Times Square' as 'LOC,' which shows stronger cross-modal feature selection.

In Case 2, CoAtt-NER correctly identifies the 'Oklahoma' entity as 'ORG,' while other methods incorrectly identify it. The reason may be that for the multimodal methods PCEN and M3S, since the visual modality mainly contains information about 'PERSON', they are unable to determine whether the image information is relevant to the entity or not, and therefore they indiscriminately incorporate the visual information into the textual representation, incorrectly recognizing 'ORG' as 'PER.' The CoAtt-NER model employs cross-modal fusion based on a collaborative cross-attention system, so it can incorporate visual information into textual representations based on relevance. Since visual information is mainly related to 'PERSON,' the CoAtt-NER model incorporates it into the 'Bob Stoops' entity representation instead of 'Oklahoma' and therefore correctly recognizes them.

This case study shows that M3S and PCEN have a misclassification phenomenon in processing character-related visual information in a MNER task, whereas CoAtt-NER effectively filters irrelevant modal information through the synergistic cross-attention mechanism, which improves the accuracy of NER.

Training Time Comparison

In this paper, we compare the training time of different models to complete one iteration on the Twitter-2015 and Twitter-2017 datasets in the same hardware environment, and the results are displayed in Table 8.

Table 8. Training time to complete one iteration for different models

Method	Twitter-2015(s)	Twitter-2017(s)
GDN-CMCF	68	51
PCEN	85	77
M3S	82	73
TISGF	71	62
CoAtt-NER	97	82

From the perspective of model architecture and computational complexity, the relatively long training time of CoAtt-NER (97s for the Twitter-2015 dataset and 82s for the Twitter-2017 dataset) is mainly attributed to the inter-modal synergistic cross-attention mechanism it adopts. This mechanism dynamically computes attention weights across modalities to enhance deep inter-modal interactions, while this approach requires more computational resources to complete the learning of feature interactions compared to direct splicing or independent modeling approaches.

Compared with GDN-CMCF (68s/51s), the latter uses a gated de-entanglement module to perform only the separation of modal features, which reduces the computational effort for cross-modal information interactions, and thus the training time is shorter. PCEN (85s/77s) uses unsupervised clustering to generate visual features and introduces an inconsistent loss for entity-type matching, a process that involves more clustering computations, resulting in a longer training time. M3S (82s/73s) mainly relies on multi-granularity visual information modeling, which is more computationally intensive but still slightly lower than CoAtt-NER. TISGF (71s/62s) uses two independent scene graphs to model and then fuse them, and compared to CoAtt-NER, its textual and visual features are modeled separately, which reduces the computational burden of direct inter-modal interactions and the training time is shorter.

The increase in training time of CoAtt-NER is due to its finer cross-modal interaction mechanism, especially the dynamic adjustment of feature weights between modalities by CoAtt-NER, which improves the accuracy of feature fusion, while ALBERT combines with KG to further enhance the text feature extraction capability. These designs enhance the expressive ability and entity recognition performance of the model, and despite the relatively longer training time, more significant improvements in accuracy and generalization are obtained, making it a more suitable model for high-precision MNER.

CONCLUSION

Focusing on the existing MNER methods, there are problems such as insufficient unimodal feature extraction and insufficient text and image feature fusion. A MNER model (CoAtt-NER) is suggested in this paper, based on the cooperative attention mechanism, combining ALBERT and CLIP-ViT for KG enhancement to optimize modal feature extraction and incorporating the cooperative attention mechanism to dynamically capture deep inter-modal interactions. Experimental results show that CoAtt-NER can achieve F1 values of 76.25% and 87.31% on the Twitter-2015 and the Twitter-2017 datasets. Compared with several other advanced MNER is improved, which can enhance the accuracy of named entities. This research effectively addresses the shortcomings of traditional methods in unimodal feature extraction and modal fusion and promotes the research progress of MNER. Its impact is not limited to NER; it also provides new research directions for other multimodal tasks such as sentiment analysis and fake news detection. In the future, CoAtt-NER can further optimize the computational efficiency to adapt to large-scale datasets while exploring finer modal alignment

strategies to improve the information complementarity between different modalities. In addition, the model can be extended to medical, financial, social media, and other fields to facilitate the application of cross-domain multimodal entity recognition on the ground.

In this paper, there are still some shortcomings and deficiencies in the research of MNER, which need to be improved. Future related work will focus on the following aspects for improvement: first, the model combines two complex deep learning models, ALBERT and CLIP-ViT, which leads to a high demand for computational resources. Therefore, the introduction of more efficient lightweight models or model distillation techniques (e.g., TinyBERT, DistilCLIP, etc.) is considered to reduce the model size, lower the computational overhead, and enhance the real-time inference capability. Meanwhile, the current model mainly uses the Twitter dataset. While the use of the Twitter dataset can demonstrate the performance of the model on social media platforms, the introduction of more types of multimodal datasets is essential in order to enhance the model's generalization ability and cross-domain adaptability. In addition, considering the diversity of global social platforms and validating the cross-linguistic NER capability using multilingual datasets will further enhance the usefulness of the model.

COMPETING INTERESTS STATEMENT

The authors of this publication declare there are no competing interests.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the authors of the article.

PROCESS DATES

Received: 02/07/2025, Revision: 03/12/2025, Accepted: 03/12/2025

CORRESPONDING AUTHOR

Correspondence should be addressed to Yulan Wen; wen2025@gxuwz.edu.cn

REFERENCES

Chen, J., Lu, Y., Lin, H., Lou, J., Jia, W., Dai, D., Wu, H., Cao, B., Han, X., & Sun, L. (2023). Learning in-context learning for named entity recognition. *Annual Meeting of the Association for Computational Linguistics.*

Cheng, J., Long, K., Zhang, S., Zhang, T., Ma, L., Cheng, S., & Guo, Y. (2024). Text-image scene graph fusion for multimodal named entity recognition. *IEEE Transactions on Artificial Intelligence*, *5*(6), 2828–2839. DOI: 10.1109/TAI.2023.3326416

Gao, L. (2024). MFE-transformer: Adaptive English text named entity recognition method based on multi-feature extraction and transformer. *Computer Science and Information Systems*, *21*, 1865–1885.

Geng, J., Zhang, C., Li, L., Yang, Q., & Zeng, D. D. (2023). PCEN: Potential correlation-enhanced network for multimodal named entity recognition. *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1–6.

Guo, A., Zhao, X., Tan, Z., & Xiao, W. (2023). MGICL: Multi-grained interaction contrastive learning for multimodal named entity recognition. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management.* DOI: 10.1145/3583780.3614967

Huang, G., He, Q., Dai, Z., Zhong, G., Yuan, X., & Pun, C. (2024). GDN-CMCF: A gated disentangled network with cross-modality consensus fusion for multimodal named entity recognition. *IEEE Transactions on Computational Social Systems*, *11*(3), 3944–3954. DOI: 10.1109/TCSS.2023.3323402

Lai, X., & Jie, Q. (2023). A named entity recognition approach for electronic medical records using BERT semantic enhancement and BiLSTM. [IJSWIS]. *International Journal on Semantic Web and Information Systems*, *19*(1), 1–14. DOI: 10.4018/IJSWIS.333711

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *ArXiv, abs/1909.11942*.

Liu, C., Yang, D., Yu, B., & Bu, L. (2024). DGHC: A hybrid algorithm for multi-modal named entity recognition using dynamic gating and correlation coefficients with visual enhancements. *IEEE Access: Practical Innovations, Open Solutions*, *12*, 69151–69162. DOI: 10.1109/ACCESS.2024.3400250

Liu, F., Ren, X., Liu, Y., Lei, K., & Sun, X. (2019). Exploring and distilling cross-modal information for image captioning. *ArXiv, abs/2002.12585*.

Liu, P., Li, H., Ren, Y., Liu, J., Si, S., Zhu, H., & Sun, L. (2023). A novel framework for multimodal named entity recognition with multi-level alignments. *ArXiv, abs/2305.08372*.

Lu, D., Neves, L., Carvalho, V. R., Zhang, N., & Ji, H. (2018). Visual attention model for name tagging in multimodal social media. *Annual Meeting of the Association for Computational Linguistics.* DOI: 10.18653/v1/P18-1185

Ma, J., Jin, W., Chen, Y., Zhang, F., Qian, S., & Qiao, Y. (2024). CLMSI: A novel image-text named entity recognition method based on contrast learning and multimodal semantic interaction. *2024 11th International Conference on Behavioural and Social Computing (BESC)*, 1–7.

Ma, Y., Liu, H., Zhang, D., Gao, C., Liu, Y., & Liu, Y. (2023). A named entity recognition method enhanced with lexicon information and text local feature. *Tehnicki vjesnik - Technical Gazette*.

Mai, W., Zhang, Z., Li, K., Xue, Y., & Li, F. (2024). Dynamic graph construction framework for multimodal named entity recognition in social media. *IEEE Transactions on Computational Social Systems*, *11*(2), 2513–2522. DOI: 10.1109/TCSS.2023.3303027

Meng, L., Qi, W., Zhou, Y., & Chen, Y. (2022). News text named entity recognition based on BI-LSTM-CRF model. *2022 41st Chinese Control Conference (CCC)*, 7217–7222.

Mo, Y., Yang, J., Liu, J., Wang, Q., Chen, R., Wang, J., & Li, Z. (2023). mCL-NER: Cross-lingual named entity recognition via multi-view contrastive learning. *AAAI Conference on Artificial Intelligence*.

Moon, S., Neves, L., & Carvalho, V. (2018). Multimodal named entity recognition for short social media posts. *arxiv preprint arxiv:1802.07862*. DOI: 10.18653/v1/N18-1078

Panga, J., Yanga, X., Qiub, X., Wanga, Z., & Huanga, T. (2024). MMAF: Masked multi-modal attention fusion to reduce bias of visual features for named entity recognition. *Data Intelligence*, *6*(4), 1114–1133. DOI: 10.3724/2096-7004.di.2024.0049

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*.

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arxiv preprint cs/0306050*.

Sun, S., Deng, M., Yu, X., & Zhao, L. (2025). HREB-CRF: Hierarchical reduced-bias EMA for Chinese named entity recognition. *arXiv*. DOI: 10.48550/arxiv.2503.01217

Tourani, A., Ejaz, S., Bavle, H., Morilla-Cabello, D., Sanchez-Lopez, J. L., & Voos, H. (2025). VS-Graphs: Integrating visual SLAM and situational graphs through multi-level scene understanding. *arXiv*.

Wang, J., Yang, Y., Liu, K., Zhu, Z., & Liu, X. (2023). M3S: Scene graph driven multi-granularity multi-task learning for multi-modal NER. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 111–120. DOI: 10.1109/TASLP.2022.3221017

Wang, X., Gui, M., Jiang, Y., Jia, Z., Bach, N., Wang, T., Huang, Z., Huang, F., & Tu, K. (2021). ITA: Image-text alignments for multi-modal named entity recognition. *ArXiv, abs/2112.06482*.

Yang, L. (2024). SAMNER: Image screening and cross-modal alignment networks for multimodal named entity recognition. *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. DOI: 10.1109/IJCNN60899.2024.10651087

Yang, T., He, Y., & Yang, N. (2022). Named entity recognition of medical text based on the deep neural network. *Journal of Healthcare Engineering*, *2022*, 1–10. DOI: 10.1155/2022/3990563 PMID: 35295179

Zhai, H., (2023). MLNet: A multi-level architecture. *Frontiers in Neurorobotics*, *17*, 1–5. DOI: 10.3389/fnbot.2023.1181143 PMID: 37408584

Zhang, Q., Fu, J., Liu, X., & Huang, X. (2018). Adaptive co-attention network for named entity recognition in tweets. *AAAI Conference on Artificial Intelligence*. DOI: 10.1609/aaai.v32i1.11962

Zhang, S., Zhang, S., He, W., & Zhang, X. (2024). A web semantic-based text analysis approach for enhancing named entity recognition using PU-learning and negative sampling. *International Journal on Semantic Web and Information Systems*, *20*(1), 1–23.

Zhang, Y., Zhou, X., Yuan, J., Wang, Z., & Pan, Y. (2024). Multimodal named entity recognition model based on cross-modal feature enhancement mechanism. *2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP)*, 36–40. DOI: 10.1109/CLNLP64123.2024.00015

Zheng, X., He, X., Ren, Y., Wang, J., & Yu, J. (2023). Owner named entity recognition in website based on multidimensional text guidance and space alignment co-attention. *Multimedia Systems*, *29*(6), 3757–3770. DOI: 10.1007/s00530-023-01170-2

Dongxiu Wang, senior experimentalist. master's degree, Graduated from East China Normal University in 2010. Worked in Guangxi University of Science and Technology. Her research interests include deep learning, data mining.

Yulan Wen, Graduated from Buriram Rajabhat University in 2024,PhD degree, lecturer, Worked in Wuzhou University, Her research interests include deep learning, educational management.

Zhengxiang Qiu, Graduated from Guangxi University of Science and Technology in 2010,Bachelor's Degree, Engineer, Worked in Guangxi Boda Software Co., Ltd., Her research interests include Deep Learning, Digital Transformation of Enterprises.