IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 17 April 2024, accepted 9 May 2024, date of publication 13 May 2024, date of current version 22 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400250

RESEARCH ARTICLE

DGHC: A Hybrid Algorithm for Multi-Modal Named Entity Recognition Using Dynamic Gating and Correlation Coefficients With Visual Enhancements

CHANG LIU[1,2], DONGSHENG YANG[1], BIHUI YU[1], AND LIPING BU[1]

[1]Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110016, China
[2]University of Chinese Academy of Sciences, Beijing 101408, China

Corresponding author: Chang Liu (liuchang183@mails.ucas.ac.cn)

This work was supported in part by the Shenyang Municipal Bureau of Science and Technology under Grant number 23-407-3-29.

ABSTRACT Multimodal named entity recognition plays a crucial role in the construction process of knowledge graphs as it directly influences the quality of entity extraction and classification, which in turn affects the overall quality of knowledge graph construction. However, most existing multimodal named entity recognition algorithms do not consider the correlation between text and images. They either use visual features of images as the attention of the text modality or fuse them with textual features. In the case of multimodal tweets containing both text and images, three categories of data can be identified based on the correlation between the two: text that is related to images, text that is partially related to images, and text that is not related to images. Using irrelevant or partially relevant image features as text cross-modal attention can result in incorrect text representation, ultimately leading to misclassification of entities and negatively impacting the model's performance. To address the problem of uncertainty or negative impact caused by the lack of relevance or partial correlation between text and images, this paper proposes a visually enhanced text representation algorithm based on a hybrid of dynamic gating and correlation coefficient. We conducted experiments on two benchmark datasets, namely Twitter-2015 and Twitter-2017. The experimental results were analyzed comprehensively to showcase the strengths of the proposed model.

INDEX TERMS Multimodal named entity recognition, visually enhanced text representation, dynamic gates, correlation coefficient calculation.

I. INTRODUCTION

Named entity recognition is a crucial and fundamental component in the construction of knowledge graphs. The quality of entity recognition directly impacts the quality of the knowledge graph construction process. Early approaches to named entity recognition focused solely on the text content of tweets. However, due to the limited length of a tweet (which cannot exceed 140 characters), relying solely on text for recognition may introduce ambiguity in word meaning and lead to category deviations in the final serialized annotation results (see Figure 1). Many tweets contain not only text but also images, and the visual information in these images is highly valuable for named entity recognition tasks. In light of the rapid advancements in multimodal learning, Moon et al. [1] introduces a novel task of multimodal named entity recognition. The included tweets incorporate visual context features from images, enhancing the capabilities of text representation and effectively addressing the issue of word polysemy ambiguity. Effectively learning visual features and combining them with text features is crucial in multi-modal named entity recognition tasks. Early work,

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

such as Moon et al. [1] was the first to integrate visual context for named entity recognition. To fully leverage the visual information in multi-modal tweets, the paper proposed a general modal attention module that selectively extracts contextual information from image modalities. Although this method effectively utilizes visual context information and enhances text representation, it fails to consider the correlation between images and text. Furthermore, multi-modal tweets often include images that are unrelated or only partially related to the text. Directly fusing text and image features in these cases introduces significant noise, leading to classification errors and reduced model performance.

Various methods of multimodal information fusion, including direct fusion (serial fusion, parallel fusion) neural networks, gating mechanisms, and attention mechanisms, are utilized. Neural networks typically have static weight learning during the information fusion process, limiting flexible adjustments. In contrast, gating mechanisms offer flexibility in controlling the fusion degree of diverse modal information, enhancing interpretability. Gating mechanisms have fewer parameters and a simpler structure compared to neural networks, leading to more efficient training. They also provide superior interpretability by clearly illustrating the fusion of different modal information, aiding in the analysis of decision-making processes and model outcomes. Moreover, gating mechanisms may exhibit enhanced generalization capabilities in processing multimodal information, enabling better adaptation to diverse datasets and tasks. Consequently, recent multimodal models have integrated gating mechanisms into their fusion modules to regulate multimodal information fusion. In 2022, Chen et al. [2] proposed dynamic gating to manage the fusion level of text and image features, effectively reducing noise from irrelevant and partially relevant images to enhance the text representation. However, this approach's heavy reliance on gating can lead to imprecise and biased correspondences between fine-grained semantic units of text and images. To address this issue, Liu et al. [3] introduces a dynamic correlation matrix calculation algorithm to assess the correlation between images and text, facilitating an end-to-end dynamic measurement of their correlation.

Drawing inspiration from Chen et al. [2] and Liu et al. [3], we propose a fusion mechanism that combines dynamic correlation coefficient and dynamic gating visual guidance. In this mechanism, object-level features and whole picture features are utilized as the prefix for each self-attention layer of BERT. This approach aims to address both the issue of model over-reliance on gated units and the fine-grained semantic correspondence deviation between text and images. To reduce the bias introduced by irrelevant images on text representation, we introduce a correlation coefficient matrix calculation method for text-image pairs. Furthermore, a dynamic gate is designed for each self-attention layer of BERT to aggregate text-related image features, enhancing overall text representation. This design mitigates the negative impact caused by text-irrelevant and text-partially related images on the final text representation. In summary, the main contributions of our paper can be summarized as follows: To address these challenges, our research proposes a fusion mechanism that incorporates both dynamic correlation coefficient and dynamic gating visual guidance. This mechanism utilizes object-level features and whole picture features as the prefix for each self-attention layer of BERT. This approach not only addresses the model's over-reliance on gated units but also effectively resolves the issue of fine-grained semantic correspondence deviation between text and images. To reduce the bias introduced by irrelevant images, we design a correlation coefficient matrix calculation method for text and images. Additionally, a dynamic gate is introduced for each self-attention layer of BERT, aggregating text-related image features to enhance text representation and mitigating the negative impact of text-irrelevant and text-partially related images on the final text representation.

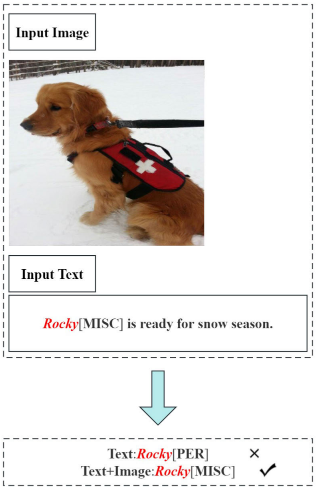

FIGURE 1. **The example of images that are irrelevant to the text.**

Furthermore, the main contributions of our paper can be summarized as follows:

(1) We utilize the FasterRCNN target detection pre-training model and the visual_grounding target detection tool to detect targets in the image datasets contained in the tweets. This step enhances the data on the image datasets within the existing datasets.

(2) We merge the strengths of CLIP for global visual feature extraction and BERT for text representation, creating a synergy between visual and textual data. This enables a more comprehensive understanding of the tweet's context.

(3) We introduce a novel method for calculating correlation scores, which evaluates the compatibility between text and image features. This innovative approach effectively weighs the relevance of visual elements to the text.

(4) Our technique integrates a dynamic gating mechanism, inspired by self-attention, that adaptively controls the

impact of visual features on the text representation, enhancing the model's capacity to filter out irrelevant information.

(5) We conduct extensive experiments to validate the effectiveness and superiority of our model. We evaluate our approach on two benchmark datasets specifically designed for multi-modal named entity recognition tasks.

II. RELATED WORK

Since 2018, the task of multimodal named entity recognition has gained significant attention in the academic community. Studies such as Moon et al. [1] introduced this task and developed the Snap Captions dataset for evaluation. Additionally, Moon et al. [4] and Zhang et al. [5] created the Twitter-2015 and Twitter-2017 datasets respectively, focusing on named entity recognition in tweets by harnessing the diverse information contained in them. Multimodal named entity recognition aims to complement the limitations of text-based approaches by incorporating visual information, thereby reducing the error rate of entity recognition. As both single-modal and multimodal learning methods advance, techniques for multimodal named entity recognition have diversified. Early approaches, like Moon et al. [1], Moon et al. [4] and Zhang et al. [5] simply and roughly fuse text and image features, and do not fully utilize the fine-grained correspondence between semantic units of different modalities. In 2021, Zhang et al. [6] proposed a Unified Multi-modal Graph Fusion (UMGF) method for Named Entity Recognition (MNER). This method utilizes a unified multi-modal graph to represent input sentences and images, where the multi-modal graph captures various semantic relationships between multi-modal semantic units (words and visual objects). Subsequently, multiple graph-based multi-modal fusion layers are stacked to iteratively perform semantic interactions for learning node representations. By enhancing the quality of text and image feature fusion, they improve the expressive power of the final representations for the ultimate purpose of named entity recognition.

The existing studies mentioned above primarily focus on integrating text and image features in multimodal named entity recognition tasks. However, it has been observed that multimodal tweets contain both image data related to the text and a substantial amount of data where the images are either unrelated or partially related to the text. Solely considering the fusion of image and text features without accounting for the correlation between them can result in uncertain or even detrimental effects on the learning process of multimodal models [7].

In order to mitigate the negative impact of irrelevant or partially relevant images on the final results, Sun et al. [7] proposes a novel pre-trained multimodal model called RIVA, based on relational reasoning and visual attention. This approach controls visual cues through gating units, effectively managing the influence of images on text semantics. Since the introduction of the RIVA model, extensive research has been conducted to investigate the correlation between text and images, with particular emphasis on selecting appropriate visual cues through gating units. For example, Sun et al. [8] proposes a multimodal BERT model for multimodal NER that selects visual cues using soft or hard gates, achieving state-of-the-art performance on public datasets. In the same year, the study by Zhao et al. [9] introduced a relation-enhanced graph convolutional network. This approach involved creating intra-modal and inter-modal relation graphs to gather image data most pertinent to the text within the dataset. By controlling the fusion level of text and image through gating mechanisms, the model selectively allowed or blocked specific information flow to improve the precision of named entity recognition, leading to cutting-edge performance at the time. Similarly, Chen et al. [2] proposes an innovative Hierarchical Visual Prefix Fusion Network (HVPNet) that leverages visual representations as insertable visual prefixes to guide text representation and decision-making. It further introduces a dynamic gating aggregation strategy for hierarchical multi-scale visual features, demonstrating superior performance on two benchmark datasets in multimodal NER.

These studies collectively demonstrate the importance of gating mechanisms in multimodal named entity recognition tasks, as they significantly improve the accuracy and efficiency of entity recognition in multimodal tweets. However, an excessive reliance on gating can introduce inaccuracies and correspondence bias between fine-grained semantic units of images and text. Some approaches have explored alternative methods that calculate correlation scores of text-image pairs to determine the fusion strength of text and image features, enhancing the final text representation. For example, Xu et al. [10] proposes a cross-modal matching module that computes similarity scores between text and images, determining the proportion of visual information to retain. Another approach, Liu et al. [3] dynamically aligns images and text sequences to facilitate multi-level cross-modal learning for enhanced word representation in text. It incorporates an end-to-end dynamic correlation measurement of image and text, eliminating the need for external tools or datasets for text-image relationship classification.

While these approaches measure or calculate correlation scores between images and text to mitigate the impact of irrelevant or partially relevant images, they do not completely resolve this issue.

In this research, we analyze the aforementioned issues and propose a solution for a multimodal named entity recognition algorithm. Our approach combines visually guided text with dynamic gating and correlation coefficients.

III. METHOD

In this section, we begin by providing the definition of the multi-modal named entity recognition task. Next, we demonstrate our model in detail, using the image and text sequence provided in Figure 2 as an example. Our proposed

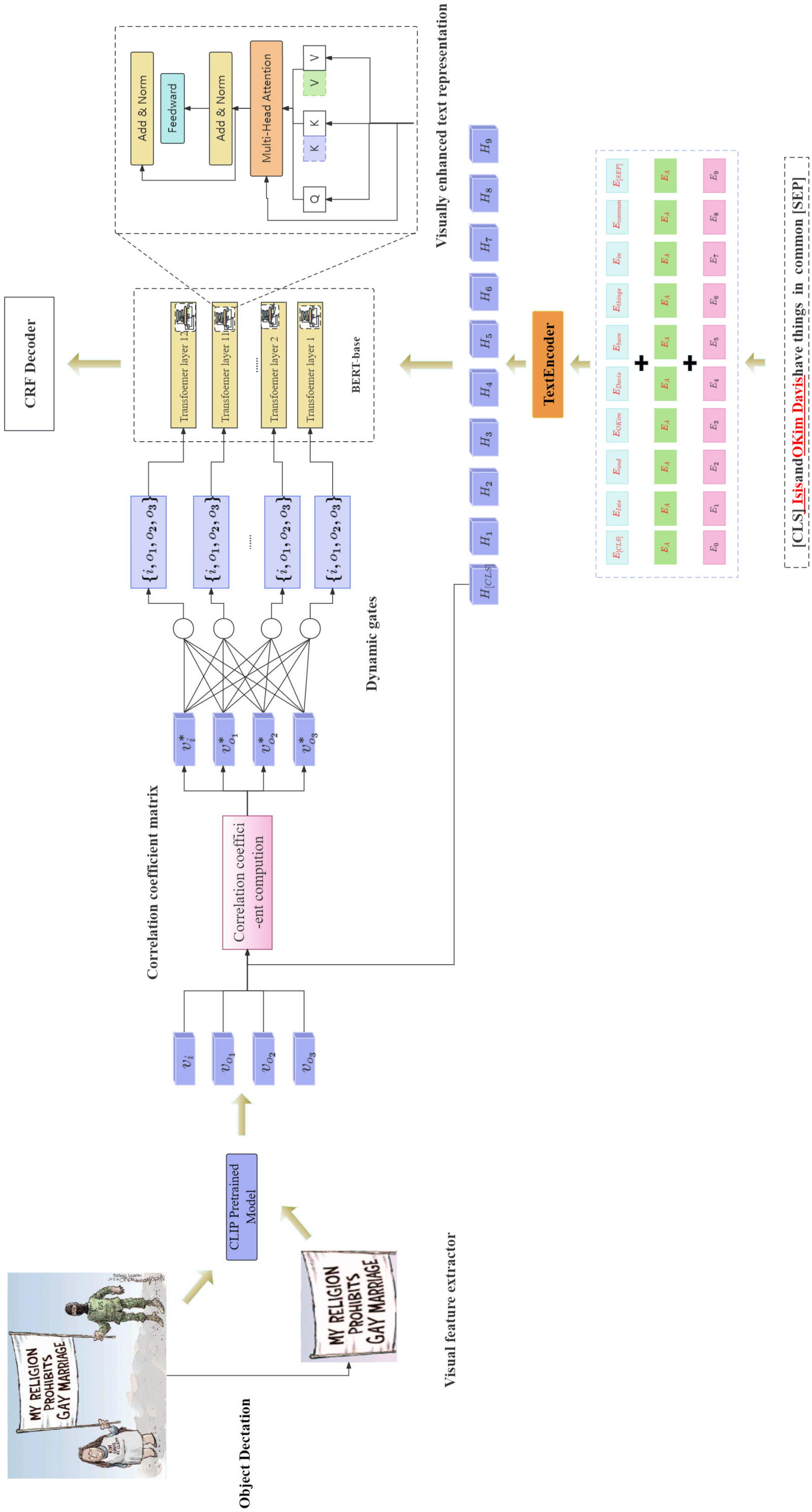
Object Dectation
MY RELIGION PROHIBITS GAY MARRIAGE
Visual feature extractor
CLIP Pretrained Model
v_i
v_{o_1}
v_{o_2}
v_{o_3}
Correlation coefficient matrix
Correlation coeffici -ent compution
v_i^*
$v_{o_1}^*$
$v_{o_2}^*$
$v_{o_3}^*$
Dynamic gates
$\{i, o_1, o_2, o_3\}$
......
Transfoemer layer 12
Transfoemer layer 11
Transfoemer layer 2
Transfoemer layer 1
BERT-base
CRF Decoder
Add & Norm
Feedward
Add & Norm
Multi-Head Attention
Q
K
K
V
V
Visually enhanced text representation
$H_{[CLS]}$ H_1 H_2 H_3 H_4 H_5 H_6 H_7 H_8 H_9
TextEncoder
$E_{[CLS]}$ E_{Isis} E_{and} E_{OKim} E_{Davis} E_{have} E_{things} E_{in} E_{common} $E_{[SEP]}$
+
E_A
+
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9
[CLS] IsisandOKim Davishave things in common [SEP]

FIGURE 2. The diagram presented above depicts the overall architecture of the model.

approach is a hybrid visual text representation enhanced through the utilization of dynamic gating and correlation coefficients for multimodal named entity recognition.In addition, the pseudocode of our model is provided on the below(Algorithm 1,Algorithm 2,Algorithm 3).

```
Algorithm $\hat{\theta} \leftarrow DGHCTraining(X_{1:N_{data}}, I_{1:N_{data}}, Y_{1:N_{data}}, \theta)$
Data: $\{(X_i, I_i, Y_i)\}_{i=1}^{N_{data}}$, a dataset of Text-Image pairs and sequence labels
Text: $X_i = \{x_1, x_2, x_3, \ldots, x_n\}$, a sequence of the $i$-th text.;
Image: $I_i = \{i, o_1, o_2, o_3\}$, original image and three objects of the $i$-th image.;
labels: $Y_i = \{y_1, y_2, \ldots, y_n\}$, a label sequence of the $i$-th text.;
Input: $\theta$, initial HMNetMNERModel parameters.;
Output: $\hat{\theta}$, the trained parameters.;
Hyperparameters: num_epochs $\in N$,$\eta \in (0, \infty)$;
for epoch = 1,2,..,num_epochs do
    for $i = 1, 2, \ldots, N_{data}$ do
        $P(\theta) \leftarrow HMNetNERModel(X_i, I_i|\theta)$;
        $loss(\theta) = -\sum_t logP\left(Y_i \middle| H^L; \theta\right)$;
        $\theta \leftarrow \theta - \eta \cdot \Delta loss(\theta)$
    end;
end
return $\hat{\theta} = \theta$
```

```
Algorithm HMNetMNERModel
Data: $\{(X_i, I_i, Y_i)\}_{i=1}^{N_{data}}$, a dataset of Text-Image pairs and sequence labels
Text: $X_i = \{x_1, x_2, x_3, \ldots, x_n\}$, a sequence of the $i$-th text.;
Image: $I_i = \{i, o_1, o_2, o_3\}$, original image and three objects of the $i$-th image.;
labels: $Y_i = \{y_1, y_2, \ldots, y_n\}$, a label sequence of the $i$-th text.;
Input: $\theta^{BERT}$, initial BERT model parameters.;
$\theta^{CLIP}$, initial CLIP model parameters.;
$\theta^{CRF}$, initial CRF decoder parameters;
Output: $\hat{\theta}^{BERT}$, $\hat{\theta}^{CLIP}$ and $\hat{\theta}^{CRF}$, the trained parameters; And P($\hat{Y}_i$),the probability of CRF classification;
Hyperparameters:prompt_dim,prompt_len,$\eta_1$, $\eta_2$ and $\eta_3 \in (0, \infty)$;
vision_embeddings $\leftarrow CLIPVisionModel(I_i, \theta^{CLIP})$;
bert_out $\leftarrow$ BERT$Model(X_i, vision\_embeddings, \theta^{BERT})$;
P($\hat{Y}_i$) = CRF(bert_out,$\theta^{CRF}$);
$loss(\theta^{BERT}, \theta^{CLIP}, \theta^{CRF}) = -\sum_t logP\left(Y_i \middle| H^L; \theta\right)$;
$\theta^{BERT} \leftarrow \theta^{BERT} - \eta_1 \cdot \Delta loss(\theta^{BERT}, \theta^{CLIP}, \theta^{CRF})$;
$\theta^{CLIP} \leftarrow \theta^{CLIP} - \eta_2 \cdot \Delta loss(\theta^{BERT}, \theta^{CLIP}, \theta^{CRF})$;
$\theta^{CRF} \leftarrow \theta^{CRF} - \eta_3 \cdot \Delta loss(\theta^{BERT}, \theta^{CLIP}, \theta^{CRF})$;
return
$\hat{\theta}^{BERT} = \theta^{BERT}$, $\hat{\theta}^{CLIP} = \theta^{CLIP}$, $\hat{\theta}^{CRF} = \theta^{CRF}$,P($\hat{Y}_i$);
end
```

A. TASK DEFINITION

The task of multi-modal named entity recognition is defined as follows: Given an input pair (X, I), where X represents a text sequence and I represents an image, the goal is to detect entities from the text sequence X and classify them according to predefined entity categories. In this study, we treat the multi-modal named entity recognition task as a serialized annotation task. We represent the input text sequence-image pair as Input $= (X = \{x_1, x_2, \ldots, x_n\}, I)$, where the text sequence contains a total of n words. Additionally, we use $Y = \{y_1, y_2, \ldots, y_n\}$ to represent the sequence labels corresponding to the text sequence. Here,$y_i \in L$, where L represents the label set in the standard BIO2 format (as described in [11]).

B. PROPOSED MODEL

In this study, we present a novel approach for multi-modal named entity recognition using a visually enhanced text representation, based on a hybrid of dynamic gating and correlation coefficients. The architecture of our model is illustrated in Figure 2, and it comprises five modules: visual feature extractor, correlation coefficient computation, dynamic gates, visual guide fusion, and multi-modal named entity recognition decoder. In the following sections, we will provide a detailed explanation of each of these modules.

C. VISUAL FEATURE EXTRACTOR

The global image features associated with sentences can express abstract concepts and have weak learning signals. The visual objects depicted in the images may be related to entities in the sentences, and these visual objects can provide additional semantics for information extraction. Therefore, we consider regional images to be important supplements of information to global images. As described in Chen et al. [2] and Zhang et al. [6], we also utilize the visual grounding toolkit, A fast and accurate one-stage approach to visual grounding, to extract the top m significant visual objects contained in the images in the multimodal tweets. This process allows us to obtain the visual objects $O = o_1, o_2, \ldots, o_m$ for image i.

In the context of the image feature extraction module, pre-trained models for image feature extraction play an important role in the field of computer vision, as they can help accelerate model training and improve model performance. Therefore, this article uses pre-trained models for image feature extraction to perform feature extraction. The

Algorithm BERT Model With Correlation Coefficient Matrix And Dynamic Gates

```
**Data**: $\{(X_i, V_i)\}_{i=1}^{N_{data}}$, a set of text sequence and
    vision_embeddings;
Text: $X_i = \{x_1, x_2, x_3, \ldots, x_n\}$, a sequence of the $i$-th
    text.;
vision_embeddings: $V_i = \{v_i, v_{o_1}, v_{o_2}, v_{o_3}\}$;
$W_{TI}$, a parameter matrix;
$W_I$, a parameter matrix;
$W_T$, a parameter matrix;
gates, 12 dynamic gates of 12-layers BERT;
**Output:**$H$,visually enhanced text representation
M $\leftarrow tanh(h_{[CLS]}^T W_{TI} v_i)$;
C $\leftarrow tanh(W_T h_{[CLS]}^T + W_I v_i M)$;
$v_i^C \leftarrow Linear(Cv_i; \theta)$;
$V^C \leftarrow [v_i, v_{o_1}, v_{o_2}, v_{o_3}] * v_i^C$;
$V_g^{(l)} \leftarrow Softmax(Leaky\_ReLU(W^l(\frac{1}{c}\sum_{i=1}^{c} v_i))), v_i \in$
    $V^C$ and $l \in [1, 12]$;
$Visual_{Attemtion}^l \leftarrow$
    $Softmax(\frac{Q^l[\phi_k^l; K^l]^T}{\sqrt{d}})[\phi_V^l; V^l], \phi_k, \phi_V = V_g^l W_l^\phi$ The
    visual prefix $V_g$ is appended to the text sequence at
    each self-attention layer of BERT.;
$H \leftarrow BERT(X_i, V_g, \theta^{BERT})$;
**return** $H$
```

pre-trained models for image feature extraction include VGG [12] proposed by the University of Oxford team, ResNet [13] proposed by Microsoft Research Institute, Inception [14] and MobileNet [15] proposed by Google Research Team, DenseNet [16] proposed by Cornell University, and a recent pre-trained model that can handle both images and text simultaneously proposed by OpenAI—CLIP [17]. Compared to traditional image feature extraction models (VGG, ResNet, Inception, MobileNet, DenseNet), CLIP can learn without annotated data, enabling the model to generalize to new categories and tasks through adversarial learning. CLIP is pre-trained on a dataset that combines over 4 billion text snippets and approximately 400,000 hours of images, allowing the model to learn from a wider and more diverse range of visual and textual information. It has excellent visual feature representation capabilities, can learn more universal feature representations, and is beneficial for achieving good performance in image feature extraction tasks. Therefore, in this work, CLIP pre-trained model is chosen as the visual feature extractor.

The input in this study consists of a text sequence-image pair, denoted as $(X, I = \{i, O\})$. Here, $X = \{x_1, x_2, \ldots, x_n\}$ represents the text sequence, where n is the length of X. And I represents a set of images, with i being the global image. For this study, we randomly select three visual object images: $O = \{o_1, o_2, o_3\}$. By feeding I as the input into the CLIP visual feature extractor, we obtain the visual feature $V = \{v_i, v_o\}$, where v_i corresponds to the global visual feature of image i, and $v_o = \{v_{o_1}, v_{o_2}, v_{o_3}\}$ represents the visual features of the visual objects in $O = \{o_1, o_2, o_3\}$.

$$V = CLIP\left(i, o_1, o_2, o_3; \theta^{CLIP}\right) \tag{1}$$

where θ^{CLIP} is the model parameter of model CLIP, $V = \{v_i, v_{o_1}, v_{o_2}, v_{o_3}\}$.

D. CORRELATION COEFFICIENT MATRIX

Unlike the approaches that rely on external tools and dataset methods, in this study, we propose a novel method that involves calculating the correlation coefficient matrix between the text global feature $h_{[CLS]}^T$ and the image global feature v_i. This calculation allows us to dynamically assess the similarity score between the image and the text. The calculation is expressed as follows:

$$M = \tanh(h_{[CLS]}^T W_{TI} v_i) \tag{2}$$

$$C = \tanh(W_T h_{[CLS]}^T + W_I v_i M) \tag{3}$$

Among them, the activation function used is tanh(·), and we have learnable parameter matrices W_{TI}, W_{TI}, and W_I. To obtain the final visual feature V^C, we calculate it based on the correlation coefficient matrix C that has been obtained.

$$v_i^c = Linear(Cv_i; \theta^C) \tag{4}$$

$$V^C = \left[v_i, v_{o_1}, v_{o_2}, v_{o_3}\right] * v_i^C \tag{5}$$

In this context, the function *Linear* (·) represents a linear function, with θ^C serving as its parameter. The visual feature V^C is defined as $[v_i^*, v_{o_1}^*, v_{o_2}^*, v_{o_3}^*]$.

E. DYNAMIC GATES

The final visual feature $V^C = [v_i^*, v_{o_1}^*, v_{o_2}^*, v_{o_3}^*]$ is divided into 4 visual blocks, with each image represented as a separate visual block. Dynamic gates are utilized to predict a normalized vector that determines the performance of each visual block. In the dynamic gate of this project, $gates_i^{(l)} \in [0, 1]$ represents the path probability from the i-th visual block to the i-th layer of Transformers in BERT. The calculation method for dynamic gates is as follows:

$$gates^{(l)} = Gates^{(l)}(V^C) \tag{6}$$

where $Gates^{(l)}(\cdot)$ represents the dynamic gate function of the l-th layer of Transformers. First, we generate the gating signal $a^{(l)}$ as follows:

$$a^{(l)} = Leaky\_ReLU(W_l(\frac{1}{c}\sum_{i=1}^{c} v_i)) \qquad v_i \in V^C \tag{7}$$

Among them, the $Leaky_ReLU$ (·) activation function is utilized, and the value of $c = 4$ represents the number of visual blocks.

Next, the features from the average vector generation of 4 visual blocks are passed through the dynamic gating unit. We utilize the MLP layer's W_l to reduce the feature dimension to c, and subsequently utilize the continuous value generated, $a^{(l)}$, as the path probabilistic soft gates. Finally, we generate the probability vector $gates^{(l)}$ for the l-th layer of Transformers, according to the following equation:

$$gates^{(l)} = Softmax\ (a^{(l)}) \tag{8}$$

Among them, $V_g^{(l)} = [V_g^{(l,i)}, V_g^{(l,o_1)}, V_g^{(l,o_2)}, V_g^{(l,o_3)}]$, where $l \in [1, 12]$.We utilize the final V_g as the visual prefix attention of Transformers to enhance the representation of the text modality.

F. VISUALLY ENHANCED TEXT REPRESENTATION

In the field of NLP pre-trained models for text representation play an important role in extracting semantic information from text, thereby achieving better performance in various NLP tasks. Currently popular pre-trained models for text representation include BERT [18] proposed by Google, as well as models ALBERT [19] which is based on BERT and XLNet [20], GPT [21] proposed by OpenAI, RoBERTa [22] proposed by Facebook as an improvement on BERT, and DistilBERT [23] proposed by HuggingFace. Among them, ALBERT and DistilBERT are lightweight models with lower performance than BERT; GPT is suitable for generative tasks but performs lower than BERT in named entity recognition tasks; RoBERTa and XLNet outperform BERT, but they consume more computational resources and have longer training times. BERT is pre-trained on a large-scale external corpus, enabling accurate understanding and processing of natural language, with powerful text feature extraction capabilities. Moreover, BERT has been trained on large-scale datasets, familiar with language models. Therefore, fine-tuning on a small-scale target training set can achieve specific tasks, which is more efficient and accurate than training from scratch. Considering the advantages and disadvantages of the aforementioned models, as well as limited computational resources and model complexity, we choose the 12-layer Transformers BERT as the text encoder for this work.

In this study, we consider a text sequence-image pair as the input, denoted as (X, I). Here, X is a sequence of text represented by $\{x_1, x_2, \ldots, x_n\}$, where n corresponds to the length of the sequence. On the other hand, I represents four images, including one global image (i) and three visual object images (o_1, o_2, o_3). These inputs are divided into the text sequence input and the image input.

After processing the images, we obtain the final visual prefix $V_g^{(l)} = \left[V_g^{(l,i)}, V_g^{(l,o_1)}, V_g^{(l,o_2)}, V_g^{(l,o_3)}\right]$ for each of the l layers ($l \in [1, 12]$). For the text sequence side, we feed the sequence X into a 12-layer Transformers BERT to obtain the text sequence representation. During the preprocessing of the text sequence with BERT, we insert the token "[CLS]" at the beginning of the sentence to capture the global semantics of the entire sentence. Finally, the output of the Text Encoder is represented as $H = \left\{h_{[CLS]}, h_1, h_2, \ldots, h_n\right\}$, where $h_{[CLS]} \in \mathbb{R}^{1\times d}$ corresponds to the global feature of the text sequence, and h_i (where $i \in 1, 2, \ldots, n$) represents the word feature of the i-th word x_i in the sequence extracted by BERT.

$$h_i = BERT\left(x_i; \theta^{BERT}\right) \tag{9}$$

Among them, θ^{BERT} is the model parameter of BERT.

In this study, the correlation coefficient matrix is used to calculate the correlation between the global feature $h_{[CLS]}$ of the text sequence and the global feature v_i of the image in multimodal tweets. The resulting correlation coefficient is used to compute the text-related image features $V^C = \left[v_i^*, v_{o_1}^*, v_{o_2}^*, v_{o_3}^*\right]$. These features are then combined with dynamic gates to generate the final visual prefix $V_g^{(l)} = \left[V_g^{(l,i)}, V_g^{(l,o_1)}, V_g^{(l,o_2)}, V_g^{(l,o_3)}\right]$, with $l \in [1, 12]$. The visual prefix V_g is appended to the text sequence at each self-attention layer of BERT. The context representation of the text sequence in the l-th layer of BERT is denoted as $H^{l-1} \in \mathbb{R}^{n\times d}$. To begin, H^{l-1} is projected into the query/key/value vectors.

$$Q^l = H^{l-1}W_l^Q \tag{10}$$

$$K^l = H^{l-1}W_l^k \tag{11}$$

$$V^l = H^{l-1}W_l^V \tag{12}$$

Next, we transform $V_g^{(l)}$ into the embedding space for text representation through a linear transformation $W_l^{\phi} \in \mathbb{R}^{d\times 2\times d}$ in the attention module. The visual cues $\phi_k, \phi_V\ \mathbb{R}^{hw(m+1)\times d}$ are employed as follows:

$$\phi_k, \phi_V = V_g^l W_l^{\phi} \tag{13}$$

Among them, $hw(m+1)$ represents the length of the visual sequence, where m denotes the number of visual objects recognized by the target detection tool. In this particular study, we set $m = 3$.

In this work, the calculation of visual prefix attention is determined by the following formula:

$$Visual^l_{Attention} = Softmax(\frac{Q^l\left[\phi_k^l; K^l\right]^T}{\sqrt{d}})[\phi_V^l; V^l] \tag{14}$$

G. MULTI-MODAL NAMED ENTITY RECOGNITION DECODER

In this study, we approach multimodal named entity recognition as a serialized annotation task. We employ Conditional Random Field (CRF) as a classifier for this task due to its ability to model context dependencies and ensure global consistency in serialization annotation tasks. CRF has been shown to outperform other methods in serialization annotation tasks. For the part-of-speech tagging categories, we utilize standard BIO2 annotations [11].

To generate the probability of predicting the label sequence y, we input the last hidden states vector H^L from the final BERT output into the CRF layer.

$$p\left(y\middle|H^L; \theta^{CRF}\right) = \frac{\prod_{i=1}^{n} F_i\left(y_{i-1}, y_i, H^L; \theta^{CRF}\right)}{\sum_{y'\in Y}\prod_{i=1}^{n} F_i\left(y'_{i-1}, y'_i, H^L; \theta^{CRF}\right)} \tag{15}$$

TABLE 1. Basic statistical information of Twitter-2015 and Twitter-2017 datasets.

	Twitter-2015	Twitter-2017
number of entity	12800	8724
number of tweets	8257	4819
train	4000	3373
validate	1000	723
test	3257	723

TABLE 2. Detail information of Twitter-2015 and Twitter-2017 datasets.

Ent Type	Twitter-2015			Twitter-2017		
	Train	Val	Test	Train	Val	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	6176	1546	5078	6049	1234	1351
Tweets	4000	1000	3257	3373	723	723

Among these components, $F(\cdot)$ represents the potential function, Y denotes the predefined label set of the standard BIO2 labeling model, and θ^{CRF} represents the parameters of the CRF model. The model is trained by maximizing the conditional likelihood estimate of the training set:

$$Loss = \sum_{t} logp\left(Y \middle| H^{L}; \theta^{CRF}\right) \tag{16}$$

During the decoding stage, the label sequence y^* is predicted for a given input X by maximizing the following score.

$$y^* = argmax_{y \in \mathbb{Y}} p\left(y|H; \theta^{CRF}\right) \tag{17}$$

IV. EXPERIMENTS

In the following chapter, we perform experiments on two benchmark datasets, Twitter-2015 and Twitter-2017, to validate our proposed approach for multi-modal named entity recognition. Our model utilizes CLIP-base as the visual encoder and BERT-base as the text encoder. The experimental results demonstrate that our proposed method leveraging dynamic gate and correlation coefficient hybrid visually enhanced text representation outperforms other models, including both uni-modal and multi-modal approaches. uni-modal methods solely rely on textual input, whereas multimodal methods utilize both textual and visual inputs.

A. DATASET

We utilized the Twitter-2015 dataset created from Twitter tweets by Lu et al. [24] and the Twitter-2017 dataset constructed by Zhang et al. [5] to evaluate our proposed model. The specifications of the Twitter-2015 and Twitter-2017 datasets are presented in Table 1 and Table 2, respectively. The Twitter-2015 dataset comprises 12,800 entities and 8,257 tweets, which are divided into a training set of 4,000 tweets, a validation set of 1,000 tweets, and a test set of 3,257 tweets. On the other hand, the Twitter-2017 dataset consists of 8,724 entities and 4,819 tweets, which are distributed into a training set of 3,373 tweets, a validation set of 723 tweets, and a test set of 723 tweets.

B. BASELINES

We compared our model against several classic named entity recognition models, encompassing both uni-modal and multimodal approaches. For the uni-modal models, we included the BiLSTM-CRF [25], CNN-BiLSTM-CRF [26], which extended the BiLSTM-CRF by incorporating character-level word representations learned through CNN. Additionally, HBiLSTM-CRF [27], which is similar to CNN-BiLSTM-CRF but employs LSTM for character-level word feature learning. We also considered BERT [18] and BERT-CRF. As for multimodal models, we selected two types: those that do not account for the correlation between text and image, and those that do. Among the methods that disregard the correlation are UMT [28], which directly merges extracted text and visual features to facilitate pure text named entity recognition, and UMGF [6], which incorporates visual entity attributes from images into the original image features and fuses text and visual characteristics using a cross-model gate module. ITA [29]aligns images with text by means of a contrastive learning module, maps image visual features to the text feature space, and resolves the challenge of modeling image-text interaction. Conversely, MGCMT [30] achieves multi-level semantic alignment of text and images using Transformer. Considering the models that consider the correlation between text and images, HVPNet [2] controls the fusion degree of text and image information through dynamic gating and visually guides the final text representation. MAF [10] calculates the similarity score between text and images using a cross-modal matching module and utilizes this score to determine the proportion of image information to retain. Lastly, HamLearning [3] dynamically measures the correlation between image and text throughout the end-to-end process without employing external tools or datasets for additional text-image relationship classification.

In addition,though our research does not directly compare with SOTA MNER models like PGIM [31], which rely on a single large-scale model, our approach differs in that it is based on multi-modal learning, offering a novel perspective. This perspective emphasizes enhancing the depth and accuracy of named entity recognition by integrating information from various modalities such as text and images. Unlike PGIM's strategy, our method leverages the complementary nature of multi-modal data, particularly in handling complex scenarios and context-related tasks, showcasing unique strengths.

Our objective is to investigate how multi-modal fusion can enhance a model's generalization and recognition capabilities, rather than solely aiming for higher precision. Due to resource constraints and differences in experimental design, our study might not have employed datasets and computational capabilities on par with PGIM. However, within the constraints, we have achieved meaningful progress, which holds research value in itself.

Our primary goal is to delve deeper into the impact of multi-modality on named entity recognition, rather than competing directly with SOTA. We hope that our findings will provide a new theoretical framework and practical strategies for future research.

C. PARAMETERS SETTINGS

In this study, we employed fixed values for various parameters. The batch size was set to 16, the learning rate to 3e-5, warmup_ratio to 0.01, prompt_len to 4, prompt_dim to 800, seed to 1234, max_seq to 12, and dropout to 0.1. However, some parameters differed based on the specific datasets. For the Twitter-2015 dataset, the model was trained for 30 epochs, and the verification process commenced in the third epoch, as indicated by the eval_begin_epoch value of 3. Conversely, for the Twitter-2017 dataset, the model was trained for 35 epochs, and verification began in the first epoch with an eval_begin_epoch value of 1.

D. METRICS

In line with the works, Liu et al. [3] and Chen et al. [2], we assess the effectiveness of our model by measuring the single-type F1 score, overall precision, overall recall, and overall F1 score. These evaluation metrics have been extensively employed in recent studies focused on multi-modal named entity recognition.

E. MAIN RESULTS

In the experiment, in order to ensure fairness, we considered the baseline results presented in two papers: Liu et al. [3] and Chen et al. [2] The results of all comparisons are depicted in Table 3 for the overall evaluation model and Table 4 for the single type evaluation model.

(1) Based on the results from Tables 3 and 4, it is evident that BERT-CRF and BERT, using the pre-trained model, outperform CNN-BiLSTM-CRF, BiLSTM-CRF, and HBiLSTM-CRF, which utilize the non-pre-trained model BiLSTM. This indicates that the inclusion of a significant amount of external knowledge from the pre-trained model is effective.

(2) Overall, considering the results from Tables 3 and 4, the multi-modal model demonstrates superior performance compared to the uni-modal model. This highlights the effectiveness of incorporating image information as a supplement to text information in enhancing model performance and reducing the error rate in recognizing entities.

(3) In terms of multi-modal models, a comparison of the overall results reveals that the performance of image and text feature fusion methods (HamLearning, ourmodel, HVPNet, MAF), after considering the correlation between images and text, outperforms the methods (UMT, UMGF, ITA, MGCMT) that directly fuse image and text features without considering their correlation. This suggests that models that consider the correlation between images and text before fusion can effectively address the uncertainty or negative impact of images that may be irrelevant or partially relevant to the text. Consequently, these models demonstrate improved performance.

(4) Among the multi-modal models, HamLearning, ourmodel, HVPNet, MAF, and ITA are the ones that consider the correlation between images and text. To address the problem of irrelevance or partial correlation between images and text, three methods are employed: calculating the weight of the correlation score between text and image to determine the degree of text-image fusion (HamLearning, MAF), using dynamic gating to determine the fusion degree of image and text features (HVPNet), and proposing a hybrid visually enhanced text representation method (ourmodel) based on dynamic gating and correlation scores. From the results in Tables 3 and 4, it can be observed that the methods that calculate the weight of correlation scores between text and image, determining the fusion degree of text and image features, as well as our proposed hybrid method, exhibit significantly better model performance compared to the models that use dynamic gating to determine the fusion degree of image and text features. This emphasizes the crucial role of calculating text-image correlation scores in the context of multi-modal named entity recognition.

(5) Our proposed model shows differing performance on the two benchmark datasets. Table 3illustrates that in the Twitter-2017 dataset, our model achieves optimal overall accuracy and overall F1 scores, with slightly lower overall recall compared to HVPNet. However, our model exhibits poor performance in the Twitter-2015 dataset in terms of overall accuracy, overall recall, and overall F1 scores. Analyzing Table 4 reveals that our model achieves optimal performance for the PER and LOC types in terms of single F1 score in the Twitter-2017 dataset, while in the Twitter-2015 dataset, it achieves optimal performance for the PER type. Notably, in both benchmark datasets, our model consistently achieves the best performance for PER type recognition.

F. ABLATION STUDY

In order to evaluate the impact of different modules in our model, we conducted ablation experiments. For convenience, we use the following abbreviations: RE Score represents the correlation coefficient matrix module, Gating represents dynamic gates, and Augmentation represents image enhancement using the visual grounding tool and the Faster RCNN pre-trained model for target detection in multi-modal tweets. Table 5 presents the comparison between the complete model and the ablation methods, revealing the following findings:

(1) All three modules contribute to the optimal performance of the model, and removing any of these modules leads to a reduction in model performance.

TABLE 3. Performance comparison of different competitive uni-modal and multi-modal approaches.

Modality	Methods	Twitter-2015			Twitter-2017		
		Precision	Recall	F1	Precision	Recall	F1
Text	BiLSTM-CRF	68.14	61.09	64.42	79.42	73.43	76.31
	CNN-BiLSTM-CRF	66.24	68.09	67.15	80.00	78.76	79.37
	HBiLSTM-CRF	70.32	68.05	69.17	82.69	78.16	80.37
	BERT	68.30	74.61	71.32	82.19	83.72	82.95
	BERT-CRF	69.22	74.59	71.81	83.32	83.57	83.44
Text+Image	UMT	71.67	75.23	73.41	85.28	85.34	85.31
	UMGF	74.49	75.21	74.85	86.54	84.50	85.51
	HVPNet	73.87	76.82	75.32	85.84	**87.93**	86.87
	MAF	71.86	75.10	73.42	86.13	86.38	86.25
	ITA	-	-	75.60	-	-	85.72
	MGCMT	73.57	75.59	74.57	86.03	86.16	86.09
	HamLearning	77.25	75.75	**76.49**	86.99	87.28	87.13
	OurModel	72.96	74.93	73.93	**87.06**	87.64	**87.35**

TABLE 4. Performance comparison of different competitive uni-modal and multi-modal approaches.

Modality	Methods	Twitter-2015				Twitter-2017			
		PER	LOC	ORG	MISC	PER	LOC	ORG	MISC
Text	BiLSTM-CRF	76.77	72.56	41.33	26.80	85.12	72.68	72.50	52.56
	CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	87.99	77.44	74.02	60.82
	HBiLSTM-CRF	82.34	76.83	51.59	32.52	87.91	78.57	76.67	59.32
	BERT	84.72	79.91	58.26	38.81	90.88	84.00	79.25	61.63
	BERT-CRF	84.74	80.51	60.27	37.29	90.25	83.05	81.13	62.21
Text+Image	UMT	85.24	81.58	63.03	39.45	91.56	84.73	82.24	**70.10**
	UMGF	84.26	83.17	62.45	42.42	91.92	85.22	83.13	69.83
	HVPNet	-	-	-	-	-	-	-	-
	MAF	84.67	81.18	63.35	41.82	91.51	85.80	85.10	68.79
	ITA	85.6	**82.6**	64.4	**44.8**	91.4	84.8	84.0	68.6
	MGCMT	85.84	82.03	63.08	40.81	90.82	86.21	86.26	66.88
	HamLearning	85.28	82.84	**64.46**	42.52	91.43	86.26	**86.66**	69.17
	OurModel	**85.97**	80.94	60.09	38.93	**93.15**	**86.27**	85.53	69.31

TABLE 5. The ablation study for different module of ourmodel.

Settings	Twitter-2015			Twitter-2017		
	Precision	Recall	F1	Precision	Recall	F1
Default	**72.96**	**74.93**	**73.93**	87.06	87.64	**87.35**
w/o RE Score	72.45	73.89	73.17	84.97	87.05	86.00
w/o Gating	73.16	73.85	73.50	**87.14**	85.79	86.46
w/o Augmentation	72.71	74.64	73.66	86.70	**87.79**	87.24

(2) Removing the RE Score module prevents the model from effectively addressing the uncertainty or negative impact caused by irrelevant or partially related images on the final textual representation. Compared to the Gating module, the RE Score module plays a more crucial role in the model.

(3) Table 5 demonstrates a significant drop in accuracy on both datasets when the RE Score module is removed, with the lowest accuracy among all models. This suggests that the removal of the RE Score module introduces misleading noise from irrelevant or partially relevant image features, resulting in incorrect detection and recognition of entities.

(4) Interestingly, removing the Augmentation module does not lead to a significant decrease in recall on both datasets and, in fact, the recall is higher than that of the Default model. This may be attributed to suboptimal target detection results obtained by the visual grounding tool and the Faster RCNN pre-training model, resulting in lower data enhancement quality and lower recall compared to the removal of the Augmentation module.

G. GENERALIZATION ANALYSIS

Considering the variances in data characteristics across different datasets, we performed cross-validation on two datasets to evaluate the model's generalization ability. Table 6 illustrates the training and testing scenarios, namely Twitter2015->Twitter2017 (training on Twitter-2015 dataset and testing on Twitter-2017 dataset) and Twitter2017->Twitter2015 (training on Twitter-2017 dataset and testing on Twitter-2015 dataset). Table 6 reveals that our model surpasses other models in accuracy only in the Twitter2017->Twitter2015 experiment, with a slightly lower F1 score compared to the optimal model. However, the model performs poorly in all three metrics in the Twitter2015->Twitter2017 experiment. These findings suggest that the generalization ability of the model trained on the Twitter-2017 dataset is superior to that of the model trained on the Twitter-2015 dataset. Possible reasons: (1) Our

TABLE 6. The performance comparison of generalization ability between ourmodel and other methods. Results with * are from paper [3].

Method	Twitter2015->Twitter-2017			Twitter2015->Twitter2017		
	Precision	Recall	F1	Precision	Recall	F1
UMT*	67.80	55.23	60.87	64.67	63.59	64.13
UMGF*	69.88	56.92	62.74	67.00	62.81	66.21
HamLearning*	**71.03**	**59.40**	**64.70**	69.17	**66.84**	**67.98**
OurModel	63.16	58.62	60.81	**70.95**	62.68	66.56

model has higher complexity compared to several existing MNER models; (2) In terms of preprocessing for image data, directly using existing object detection tools resulted in poor object detection performance, leading to subpar model training performance and consequently weak generalization ability; (3) Due to time constraints, we did not select the optimal hyperparameters through methods like cross-validation, resulting in improper hyperparameter selection and causing weak generalization performance of our model. In future research, we will continue to optimize our model to enhance its generalization ability.

V. CONCLUSION AND FUTURE WORK

In this research, our proposed method for multimodal named entity recognition revolves around a hybrid approach that combines dynamic gating and correlation scores to enhance text representation through visual elements. Specifically, we employ the CLIP pre-training model to extract global visual features and object-level visual features from the entire image. In parallel, we utilize the BERT pre-training model to extract global features from the text of multi-modal tweets and individual word features. We introduce a correlation score calculation matrix to determine the correlation scores between the text and image by evaluating the global visual features of the original image and the global text features. Subsequently, we derive the final visual features based on the correlation scores. To enhance text representation further, we propose dynamic gating, which guides visual features similar to each self-attention prefix in BERT. We conducted extensive experiments, ablation analysis, and generalization analysis on two benchmark datasets. The experimental results demonstrate the effectiveness and robustness of our proposed method. The ablation analysis shows that the RE Score module can reduce uncertainties and negative impacts caused by unrelated images or partially related images to the final text representation, the Gating module can effectively control the fusion degree of text and images, and the object detection tools used in the Augmentation module do not yield ideal results. The generalization analysis indicates that our model performs poorly in terms of generalization. In future research, we will continue to optimize our model to improve its generalization capabilities.

ACKNOWLEDGMENT

The research is supported by the Shenyang City Science and Technology Plan, the project name is: "Research on Key Technologies for Self-Regulation of Digital Educational Resources Driven by Data and Knowledge". Project number: 23-407-3-29.

REFERENCES

[1] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," 2018, *arXiv:1802.07862*.

[2] X. Chen, N. Zhang, L. Li, Y. Yao, S. Deng, C. Tan, F. Huang, L. Si, and H. Chen, "Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction," 2022, *arXiv:2205.03521*.

[3] P. Liu, H. Li, Y. Ren, J. Liu, S. Si, H. Zhu, and L. Sun, "Hierarchical aligned multimodal learning for NER on tweet posts," 2023, *arXiv:2305.08372*.

[4] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," 2018, *arXiv:1802.07862*.

[5] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 5674–5681.

[6] D. Zhang, S. Wei, S. Li, H. Wu, Q. Zhu, and G. Zhou, "Multi-modal graph fusion for named entity recognition with targeted visual guidance," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 16, 2021, pp. 14347–14355.

[7] L. Sun, J. Wang, Y. Su, F. Weng, Y. Sun, Z. Zheng, and Y. Chen, "RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1852–1862.

[8] L. Sun, J. Wang, K. Zhang, Y. Su, and F. Weng, "RpBERT: A text-image relation propagation-based BERT model for multimodal NER," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 15, pp. 13860–13868.

[9] F. Zhao, C. Li, Z. Wu, S. Xing, and X. Dai, "Learning from different text-image pairs: A relation-enhanced graph convolutional network for multimodal NER," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3983–3992.

[10] B. Xu, S. Huang, C. Sha, and H. Wang, "MAF: A general matching and alignment framework for multimodal named entity recognition," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Feb. 2022, pp. 1215–1223.

[11] J. Li, A. Sun, J. Han, and C. Li, "A Survey on deep learning for named entity recognition," *IEEE Trans. On Knowl. and Data Eng.*, vol. 34, no. 1, pp. 50–70, Mar. 2020.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8748–8763.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[19] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.

[20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[21] A. Radford et al., "Improving language understanding by generative pre-training," [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[24] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018.

[25] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[26] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," 2016, *arXiv:1603.01354*.

[27] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, *arXiv:1603.01360*.

[28] J. Yu, J. Jiang, L. Yang, and R. Xia, "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3342–3352.

[29] X. Wang, M. Gui, Y. Jiang, Z. Jia, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu, "ITA: Image-text alignments for multi-modal named entity recognition," 2021, *arXiv:2112.06482*.

[30] P. Liu, G. Wang, H. Li, J. Liu, Y. Ren, H. Zhu, and L. Sun, "Multi-granularity cross-modality representation learning for named entity recognition on social media," 2022, *arXiv:2210.14163*.

[31] J. Li, H. Li, Z. Pan, D. Sun, J. Wang, W. Zhang, and G. Pan, "Prompting chatgpt in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2023.

CHANG LIU received the B.S. degree from Shenyang Aerospace University, Shenyang, in 2017. She is currently pursuing the Ph.D. degree with Shenyang Institute of Computing Technology, University of Chinese Academy of Sciences. Her research interests include knowledge engineering, natural language processing, multi-modal learning, and deep learning.

DONGSHENG YANG received the B.S. degree from Jilin University, China, in 1986. He is currently a Professor and a Ph.D. Supervisor with Shenyang Institute of Computing Technology, Chinese Academy of Sciences. His research interests include the Internet of Things, machine learning and pattern recognition, real-time control systems, and numerical control. His awards and honors include the Technology Progress of Chinese Academy of Science and Liaoning Province Technology Progress.

BIHUI YU received the B.S. and M.S. degrees from Xidian University, and the Ph.D. degree from Shenyang Institute of Computing Technology, Chinese Academy of Sciences. He is currently a Professor and a Ph.D. Supervisor with Shenyang Institute of Computing Technology. His research interests include knowledge engineering, big data technology, multimodal analysis, and semantic web.

LIPING BU received the B.S. degree from Qufu Normal University and the M.S. degree from Shandong University of Science and Technology. She is currently an Associate Research Fellow and a M.S. Supervisor with Shenyang Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include big data, artificial intelligence, and knowledge engineering.