Adaptive Global-Local Contrastive Learning and Information Enhancement for Multimodal Entity and Relation Extraction

Tiantian Chen, Meiling Wang*, Fang Yang
School of Cyberspace Security and Computer Science
Hebei University, Baoding, China
Emails: chentiantianlf@163.com, wmling16@163.com, yangfang@hbu.edu.cn

***Abstract*—Multimodal Named Entity Recognition (MNER) and Multimodal Relation Extraction (MRE) extract entities and relations from text by fusing complex visual data. Despite notable advances, existing methods suffer from two core challenges: The first issue is the key information loss, where noise suppression eliminates entity-related visual cues in MNER and relation-critical details in MRE, degrading performance; The second issue is insufficient original information, which triggers entity misjudgment in MNER and relation determination deviations in MRE. This paper proposes a multimodal joint task network (AGLCLIE) to optimize MNER and MRE simultaneously. Specifically, an adaptive global-local contrastive learning module is designed to address the issue of key information loss: it constructs a Consistency-Aware Factor to quantify the semantic consistency between the textual and visual modalities and achieves adaptive and dynamic adjustment of temperature parameters in the two dimensions of "word-image patch" (global) and "phrase-region" (local). In addition, an information enhancement module is designed to address the problem of insufficient original information: it guides the Large Language Model to generate accurate explanations of entities and nouns through prompts, so as to supplement semantics and eliminate ambiguities. Experimental results show that our model achieves F1-scores of 75.97%, 87.82% and 87.27% on the Twitter15, Twitter17 and MNRE datasets, exceeding the prior state-of-the-art methods by 0.65%, 0.95% and 4.14%, respectively. These results confirm that AGLCLIE outperforms current SOTA methods on both MNER and MRE tasks.**

***Index Terms*—Multimodal Named Entity Recognition, Multimodal Relation Extraction, Adaptive Contrastive Learning, Information Enhancement**

I. Introduction

Multimodal Named Entity Recognition (MNER) and Multimodal Relation Extraction (MRE) extend text-only NER and RE by leveraging images as auxiliary inputs to boost entity discrimination and relational reasoning [1], [2]. While early studies have advanced the joint training of these two tasks, two core challenges remain unresolved: (1) Key information loss: Both tasks require retaining critical visual cues for entities and relations (e.g., sports-related details in Fig. 1(a) and textual content in Fig. 1(b)).(2) Insufficient original information: Entities lack adequate semantic descriptions, and supplementing such gaps with noun or entity explanations can effectively improve task performance (as shown in Fig. 1(c) and Fig. 1(d)).

Fig. 1. Samples for the MNER and MRE task and issues with previous methods.

This paper proposes the Adaptive Global-Local Contrastive Learning and Information Enhancement Network (AGLCLIE). To address the issue of key information loss, the model is equipped with an adaptive global-local contrastive learning module. This module adopts a dual-channel cross-attention mechanism to extract textual and visual semantic features. Then, based on the global-local contrastive learning architecture, it calculates the cosine similarity between these two types of features to construct a Consistency-Aware Factor (serving as the temperature parameter for contrastive loss calculation), thereby adaptively adjusting the distance between feature representations. Subsequently, the module fuses global-local dual-dimensional features via a Multi-Layer Perceptron (MLP) to generate the module output. To tackle the problem of insufficient raw information, we design an information enhancement module. Leveraging a Large Language Model (LLM), this module parses the text, extracts precise explanations of entities and nouns, and incorporates these semantic details into the original textual information to enhance the capability of semantic representation. Finally, the outputs of the two modules are fused and fed into a classifier, leading to a significant improvement in the F1-score.

*Corresponding author: Meiling Wang, E-mail: wmling16@163.com.

II. Related Work

Early studies [3],[4] adopted RNNs for text encoding and CNNs for visual feature extraction, fusing full-image features with text. Subsequent research has verified the superiority of object-level feature fusion and cross-modal pre-training [5], [6]. Yu et al. [7] and Zhang et al. [8] proposed using region-based image features (e.g., object detection box features) for accurate entity representation and introduced the Transformer architecture to explore fine-grained semantic correspondence between text and vision. Chen et al. [1] presented the HVPNeT model, optimizing BERT's multimodal input mechanism via visual feature prefix injection. In terms of the application of contrastive learning, Khosla et al. [9] proposed supervised contrastive learning to enhance representation discriminability by reducing the distance between intra-class samples. Yang et al. [10] proposed the CAG framework, which adopts a dynamically adjusted contrastive strategy to enhance the consistency of multiple types of image-text pairs. In terms of the application of large models, Li et al. [11] proposed the PromptNER model, guiding Large Language Models (LLMs) to complete few-shot NER tasks via prompt engineering. He et al. [12] proposed the HGMAF model, which leverages GPT-3.5-Turbo to generate hierarchical text information (e.g., sentiment, entities) and adopts diffusion models to produce semantically rich images.

Although these methods have achieved good results, it is still difficult to balance information retention and noise filtering and address the issue of insufficient original semantic information. Thus, the method we design, which integrates solutions to these two problems, has achieved excellent performance.

III. Methodology

AGLCLIE aims to achieve precise alignment and deep integration of multimodal data via the adaptive global-local contrastive learning module and utilize external knowledge enhancement strategies to enrich the original semantic information through the knowledge enhancement module. The overall architecture of the network is illustrated in Fig. 2. In subsequent sections, each of the aforementioned functional modules will be elaborated in detail.

A. Adaptive Global-Local Contrastive Learning Module

This module employs a two-level partitioning strategy: textual input is split into word-level T_w and phrase-level T_p embeddings, both generated via BERT encoding [13]; visual input is parsed into grid-derived patch-level V_p and toolkit-extracted region-level V_r embeddings [14]. Consistent with prior work [1], [15], the numbers of image patches and regions are fixed at 49 and 3, respectively, with their visual embeddings generated by ResNet50 encoding [16]. A dual-channel cross-attention network fuses global text T_w and visual V_p embeddings, as well as local text T_p and visual V_r embeddings. The calculation of cross-attention mechanism fusion is shown in (1-7):

$$CrossAtt(\cdot) = softmax\left(\frac{(MW_Q)(NW_K)^T}{\sqrt{d_k}}\right)NW_V, \quad (1)$$

$$Concat[\cdot] = concat[CrossAtt(M,N), M], \quad (2)$$

$$gate(\cdot) = sigmoid(Concat[M,N]W_g), \quad (3)$$

$$H_T^{Global} = (gate(T_w, V_p) * Concat[T_w, V_p])W_l, \quad (4)$$

$$H_V^{Global} = (gate(V_p, T_w) * Concat[V_p, T_w])W_l, \quad (5)$$

$$H_T^{Local} = (gate(T_p, V_r) * Concat[T_p, V_r])W_l, \quad (6)$$

$$H_V^{Local} = (gate(V_r, T_p) * Concat[V_r, T_p])W_l, \quad (7)$$

where M and N respectively represent the input feature matrices of different modalities,W_Q,W_K,W_V,W_g,W_l represent the learnable linear mapping weight matrices, d denotes the feature dimension. For the functions $CrossAtt(\cdot)$, $Concat[\cdot]$ and $gate(\cdot)$, the first input parameter is M and the second input parameter is N.

For a given sample record r, the representations of different modalities from the same sample are defined as positive sample pairs, while the modal representations from other sample records r' are regarded as negative samples. An in-batch negatives mechanism is adopted in the contrastive learning process for optimization.The calculation of contrastive learning is shown in (8):

$$L_C(\cdot) = -\sum_{r\in R} log\frac{exp(H_M^r V_N^r \ / \ \tau)}{\sum_{r'\in B} exp(H_M^r V_N^{r'} \ / \ \tau)}, \quad (8)$$

where B denotes a batch of displayed data, R represents the set of positive sample pairs,and τ is the temperature parameter that controls the contrast intensity.

Leveraging the negative correlation between temperature and constraint intensity [17], [18], this paper proposes an adaptive strategy. We modify the classic contrastive learning framework [2], [19] by computing a Consistency-Aware Factor via text-image cosine similarity, which serves as a dynamic temperature parameter to quantify image-text consistency. The calculation formula of the Consistency-Aware Factor is shown in (9):

$$\triangle CA = \epsilon - \alpha * CosineSimilarity(H_M, H_N), \quad (9)$$

where H_M and H_N respectively represent fusion features of different modalities, ϵ denotes temperature baseline coefficient, α represents the dynamic adjustment coefficient.

The global-local contrastive loss is defined as shown in (10–11):

$$L^{Global} = (L_C(H_T^{Global}, H_V^{Global}) + L_C(H_V^{Global}, H_T^{Global}))/2, \quad (10)$$

$$L^{Local} = (L_C(H_T^{Local}, H_V^{Local}) + L_C(H_V^{Local}, H_T^{Local}))/2. \quad (11)$$

To achieve multi-dimensional dynamic fusion, the output features H_T^{Global}, H_V^{Global}, H_T^{Local} and H_V^{Local} obtained after backpropagation of adaptive contrastive learning are input into

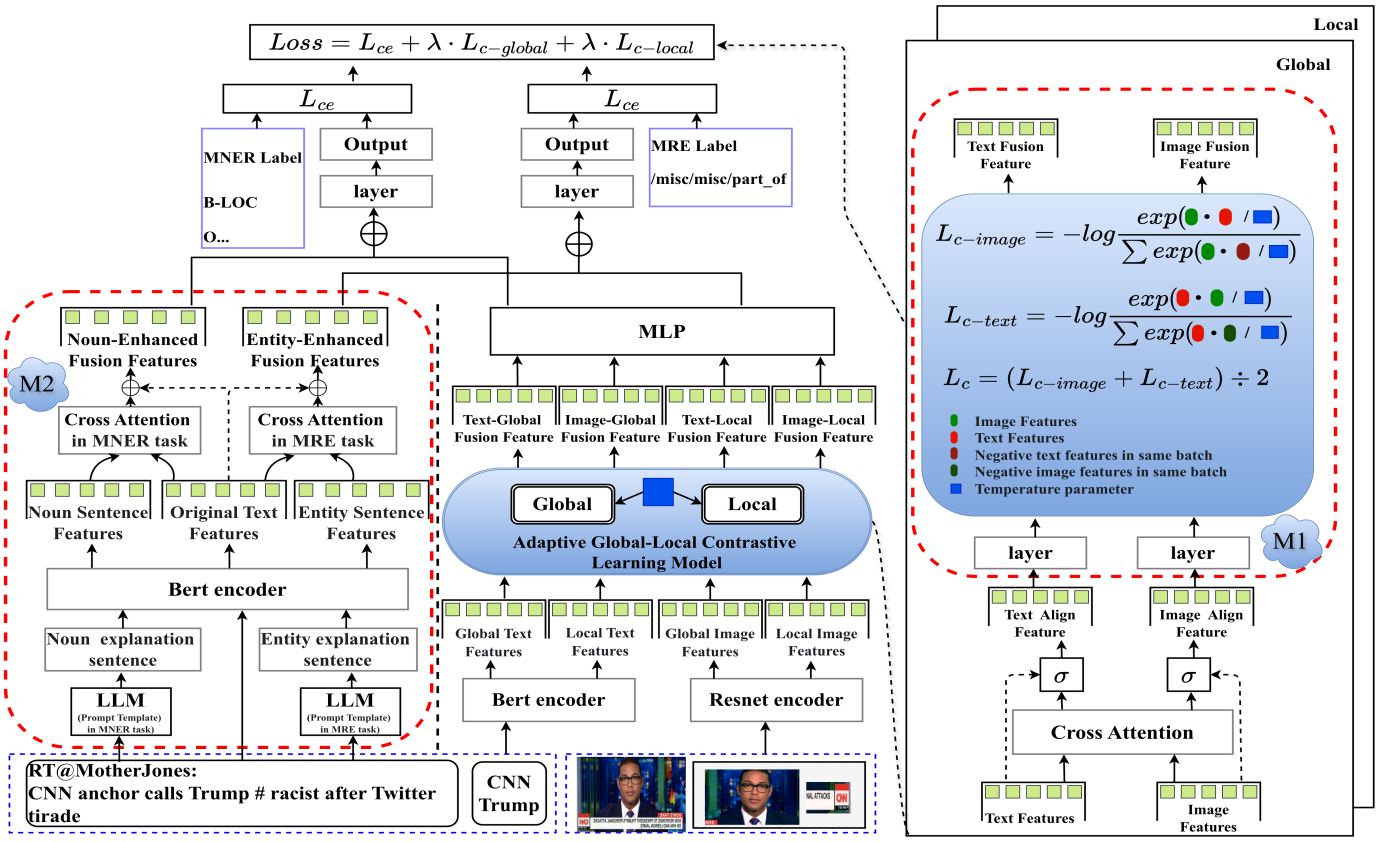

Fig. 2. The Overall Architecture of the AGLCLIE Network, M1 represents the core component of the adaptive global-local contrastive learning module, and M2 represents the information enhancement module.

MLP network module for weighted integration. The specific calculation process is shown in (12-14):

$$H_{concat} = [H_T^{Global}, H_V^{Global}, H_T^{Local}, H_V^{Local}], \quad (12)$$

$$gate_{mlp} = softmax(H_{concat} W_{mlp}), \quad (13)$$

$$output_{mlp} = gate_{mlp} * H_{concat}, \quad (14)$$

where W_{mlp} denotes the learnable weight matrix, $*$ represents the element-wise multiplication operation, and $output_{mlp}$ denotes the final output feature weighted by the MLP network module.

B. Knowledge Enhancement Module

We design entity and noun explanation prompt templates for the original text (see Fig. 3), where the noun template targets MNER tasks and the entity template targets MRE tasks. Using GPT-3.5-Turbo [20], we generate corresponding explanation information, whose definitions are given in (15-16):

$$G_{entity} = LLM(Prompt_{entity}(T)), \quad (15)$$

$$G_{noun} = LLM(Prompt_{noun}(T)), \quad (16)$$

where G_{entity} denotes MRE-related explanation information of entities, and G_{noun} denotes MNER-related explanation information of nouns.

Prompt Temple for Entity Interpretations	Based on the content of sentence, briefly explain the nouns of head_entity and tail_entity, with each explanation not exceeding 15 words. The output format is "head_entity: Interpretation of head_entity, tail_entity: Interpretation of tail_entity". The input format is "sentence:<input_text>,head_entity: <head_entity>,tail_entity:<tail_entity>".
Prompt Temple for noun Interpretations	Extract the nouns or noun phrases in the following sentence that are most likely to be entities, with a maximum of 3. The output format is "['noun 1', 'noun 2', 'noun 3']".The input format is "sentence:<input_text>". Based on the content and context of the following sentence, briefly explain the nouns in the noun list, with each explanation not exceeding 15 words. The output format is "Noun 1: Interpretation of Noun 1, Noun 2: Interpretation of Noun 2, Noun 3: Interpretation of Noun 3".The input format is "sentence: <input_text>,list of nouns:<noun_list>".

Fig. 3. GPT-3.5-Turbo Prompt Temple, ⟨⟩ represents the input variable.

Subsequently, the model employs BERT to encode T_w, G_{entity} and G_{noun}, yielding the text embedding B_T, entity explanation embedding B_{entity} and noun explanation embedding B_{noun}, respectively. Text enhancement that integrates entity and noun explanations is implemented via the cross-attention network, with the calculation formulas shown in (17–18):

$$C_{entity} = CrossAtt(B_T, B_{entity}) + B_T, \quad (17)$$

$$C_{noun} = CrossAtt(B_T, B_{noun}) + B_T. \quad (18)$$

C. Classifier

For the MRE task, $output_{mlp}$ is concatenated with C_{entity}. After FFN processing, we obtain the task output $output_{MRE}$.

The $[CLS]$ token and $softmax$ function are then used to aggregate the probability distribution over the relation label set R. Finally, we calculate the relation extraction loss with cross-entropy loss L_{re} (19–20) and derive the total loss L_{t-re} as per (21):

$$output_{MRE} = FFN(concat(output_{mlp}, C_{entity})), \quad (19)$$

$$L_{re} = -\sum_{i=1}^{n} log(softmax(output_{MRE}[CLS])), \quad (20)$$

$$L_{t-re} = L_{re} + \lambda * L^{Global} + \lambda * L^{Local}. \quad (21)$$

For the MNER task, $output_{mlp}$ is concatenated with C_{noun}, and the task output $output_{MNER}$ is generated via FFN. Following Moon et al. [21] and Yu et al. [7], $output_{MNER}$ is fed into the CRF model. For the label sequence $y = \{y_1, \ldots, y_n\}$, its probability and the MNER loss L_{ner} are defined in (22–24), and the total loss L_{t-ner} is computed as per (25):

$$output_{MNER} = FFN(concat(output_{mlp}, C_{noun})), \quad (22)$$

$$p(y|output_{MNER}) = \frac{\prod_{i=1}^{n} S_i(y_{i-1}, y_i, output_{MNER})}{\sum_{y' \in Y} \prod_{i=1}^{n} S_i(y'_{i-1}, y'_i, output_{MNER})}, \quad (23)$$

$$L_{ner} = -\sum_{i=1}^{M} log(p(y^{(i)}|output^{(i)}_{MNER})), \quad (24)$$

$$L_{t-ner} = L_{ner} + \lambda * L^{Global} + \lambda * L^{Local}. \quad (25)$$

where Y represents the predefined label set based on the BIO annotation scheme, $S(\cdot)$ denotes the potential function, and M stands for the total number of samples. Relevant details can be referred to in the work by Lample et al.[22].

IV. Experiments

A. Datasets, Evaluation Methods, and Experimental Settings

We evaluate the proposed method on three public datasets: Twitter15 and Twitter17 [23] for the MNER task, and the manually annotated MNRE [24] for the MRE task. Three hyperparameters are tuned: temperature baseline coefficient ϵ, dynamic adjustment coefficient α, and contrastive loss proportion coefficient λ. Their values are set as ϵ=0.1, α=0.001, λ=0.005 (Twitter15); ϵ=0.09, α=0.001, λ=0.005 (Twitter17); and ϵ=0.1, α=0.08, λ=0.01 (MNRE). Precision, Recall, and F1-score are used as evaluation metrics, with results detailed in subsequent sections.

B. Multimodal-Based Methods

UMT [7], UMGF [25] and MEGA [24] are based on Transformer architectures or graph neural networks (GNNs). They deeply align fine-grained visual object features with textual semantic representations, thus effectively mining cross-modal correlation information. HVPNet [1] constructs a hierarchical visual structure and integrates visual prefixes into the BERT model. MKGFormer [15] uses a hybrid Transformer fusion of visual-text knowledge, which can improve the performance of information retrieval, question answering, and recommendation tasks. MGCMT [26] is a multi-granularity cross-modal Transformer model, which can efficiently integrate multi-granularity features of text content and auxiliary images. CAG [10] innovatively introduces consistency factors and contrastive learning strategies, and combines them with task-specific instruction fine-tuning schemes to achieve joint MRE. RAP [27] integrates the relevance-aware prompt-tuning and dynamic router mechanism for image-text alignment.

C. Analysis of Comparative Results with Baseline Methods

The comparison results with baseline models are shown in Table I and Table II. For the MNER task, the proposed method achieves a performance of 87.82% on the Twitter17 dataset, which is not only better than UMGF (85.51%) and UMT (85.31%) that focus on fine-grained object-level image-text alignment, but also outperforms HVPNet (86.87%) which adopts hierarchical visual representation alignment. For the MRE task, compared with the HVPNet model adopting hierarchical visual representation, the MKGFormer model utilizing powerful vision-language pre-trained embeddings, and the RAP model employing the relevance-aware prompt-tuning and dynamic router mechanism to suppress noise, the proposed method achieves significant performance improvements (in terms of F1-score, the proposed method is 5.42, 5.32, and 4.14 percentage points higher respectively).

TABLE I
PERFORMANCE COMPARISON OF MODELS ON MNER TASKS

Model	Twitter-2015			Twitter-2017		
	Prec.	Rec.	F1	Prec.	Rec.	F1
UMT	71.67	75.23	73.41	85.28	85.34	85.31
UMGF	74.49	75.21	74.85	86.54	84.5	85.51
MEGA	70.35	74.58	72.35	84.03	84.75	84.39
HVPNeT	73.87	**76.82**	75.32	85.84	87.93	86.87
MGCMT	73.57	75.59	74.57	86.03	86.16	86.09
AGLCLIE(ours)	**75.39**	76.57	**75.97**	**86.83**	**88.82**	**87.82**
w/o M1	74.17	75.8	74.98	86.89	86.31	86.59
w/o M2	74.48	76.64	75.55	87.3	86.97	87.13
w/o AT	75.07	76.26	75.66	86.32	86.38	86.35
w/o T	74.36	74.89	74.62	85.66	86.68	86.17

TABLE II
PERFORMANCE COMPARISON OF MODELS ON MRE TASKS

Model	Prec.	Rec.	F1
UMT	62.93	63.88	63.46
UMGF	64.38	66.23	65.29
MEGA	64.51	68.44	66.41
HVPNeT	83.64	80.78	81.85
MKGFormer	82.67	81.25	81.95
CAG	82.8	81.71	82.25
RAP	83.59	82.67	83.13
AGLCLIE(ours)	**86.14**	**88.43**	**87.27**
w/o M1	80.09	82.34	81.2
w/o M2	86.35	87.03	86.69
w/o AT	82.65	85.62	84.11
w/o T	81.54	85.62	83.53

D. Analysis of Ablation Experiments

In this section, MNER task results are shown in Table I; MRE task results in Table II. M1 represents the core component of the adaptive global-local contrastive learning module, and M2 represents the information enhancement module; AT represents adaptive temperature, and T represents temperature.

We conduct comparative experiments to verify the effectiveness of each core module. Removing the M1 module causes the model's F1-score to drop by up to 6.07%, which indicates that the adaptive contrastive learning can effectively alleviate the problem of image-text mismatch. Removing the M2 module results in F1-score degradation across all three datasets, demonstrating that entities and noun explanations can provide contextual details and semantic understanding for the model.

To verify the effectiveness of the dynamic temperature adjustment mechanism for contrastive learning in the model, we design two sets of control experiments: one uses only fixed temperature parameters, and the other removes the temperature parameters entirely. Experimental results on the three datasets confirm that dynamic temperature adjustment can significantly improve the model performance.

E. Analysis of the Impact of Different LLMs on Model Performance

To investigate how different large language models (LLMs) affect target task performance, this study tests three state-of-the-art mainstream models: Qwen-Max [28], Meta-Llama-3.1-8B-Instruct [29] and GPT-3.5-Turbo [20]. Experimental results (Fig. 4) show that GPT-3.5-Turbo gets the best performance in the MRE-MNER dual-task setting. Further analysis of

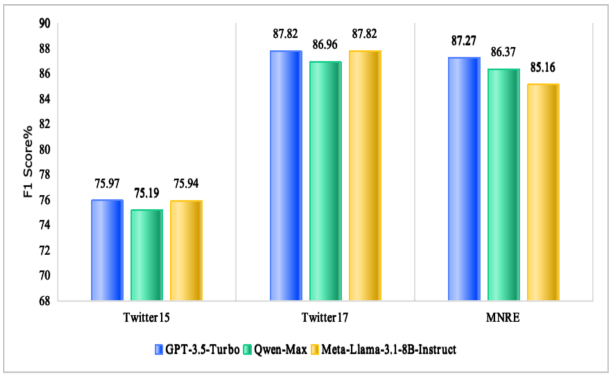

Fig. 4. The Impact of LLMs on Model Performance.

model outputs (Fig. 5) finds that even with the same prompt templates, explanations from different LLMs differ a lot in style and quality. Qwen-Max generates concise explanations for nouns and entities with low information density. Meta-Llama-3.1-8B-Instruct focuses more on complete details in its explanations. GPT-3.5-Turbo produces explanations that are more linguistically vivid. It also better integrates the context of the input text and achieves deep semantic alignment with the original text, which is a key factor supporting its superior performance in the dual tasks.

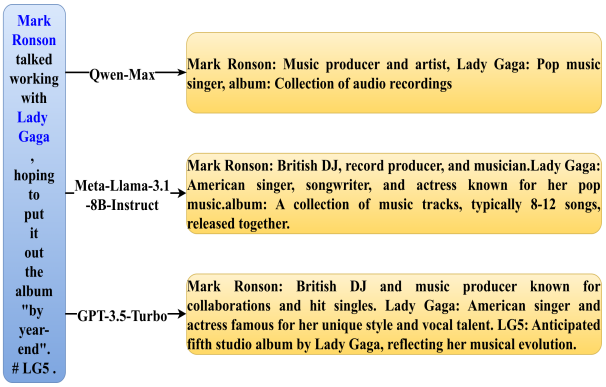

Fig. 5. The Explanation Results Generated by Different LLMs.

F. Case Analysis

We compare the performance of the HVPNeT and AGLCLIE models on typical cases of the MNER and MRE tasks. In the MNER task (as illustrated in Fig. 6(a) and Fig. 6(b)), the HVPNeT model relies on an attention mechanism for image-text alignment, and is thus prone to recognition errors in scenarios with low image-text matching and semantic ambiguity. In the MRE task (as shown in Fig. 6(c) and Fig. 6(d)), relation judgment has to rely solely on textual information when image-text matching is poor. Specifically, for the spousal relationship between Meghan and Harry, the HVPNeT model incorrectly classifies it as the "/person/person/peer" relation due to the lack of semantic information for direct relation judgment. In contrast, our method leverages the introduced external knowledge of "Meghan Markle: A prominent figure, part of the royal family, involved in official royal tours. Harry: Given name of Prince Harry, partner of Meghan Markle", enabling the model to directly determine that the relationship between these two entities is "/person/person/couple".

Fig. 6. Case Analysis for MNER and MRE Task.

V. Conclusion

This paper proposes the AGLCLIE network model, which effectively mitigates the problems of modal misalignment

and insufficient information in multimodal benchmark tests. Drawing on the core ideas of contrastive learning methods, we implemente a multimodal version of the global-local multi-dimensional adaptive contrastive learning technology, which can maximize the retention of key information across different modalities. Meanwhile, we leverage LLMs to extract explanatory information about entities and nouns, thereby enhancing the semantic information of the text. The model has demonstrated excellent performance on three core datasets, fully proving its strong capabilities in multimodal data understanding and processing tasks.

In the future, relying on the information introduction mechanism of the information enhancement module, we aim to construct a task-oriented external knowledge fusion framework by integrating more advanced encoders with next-generation large language models, thereby improving the accuracy of semantic representation for entity relations.

ACKNOWLEDGMENT

The authors acknowledge the support of GPT-3.5-Turbo(OpenAI, 2022), Meta-Llama-3.1-8B-Instruct(Meta Platforms, Inc., 2024), and Qwen-Max(Alibaba Cloud, 2023) for part of the data generation in this experiment.

REFERENCES

[1] Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., ... Chen, H. (2022). Good visual guidance makes a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. arXiv preprint arXiv:2205.03521.

[2] Hu, X., Guo, Z., Teng, Z., King, I., Yu, P. S. (2023). Multimodal relation extraction with cross-modal retrieval and synthesis. arXiv preprint arXiv:2305.16166.

[3] Arshad, O., Gallo, I., Nawaz, S., Calefati, A. (2019, September). Aiding intra-text representations with visual context for multimodal named entity recognition. In 2019 International conference on document analysis and recognition (ICDAR) (pp. 337-342). IEEE.

[4] Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H. (2018, July). Visual attention model for name tagging in multimodal social media. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1990-1999).

[5] Hu, X., Chen, J., Liu, A., Meng, S., Wen, L., Yu, P. S. (2023, October). Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In Proceedings of the 31st ACM international conference on multimedia (pp. 5185-5194).

[6] Sun, L., Wang, J., Zhang, K., Su, Y., Weng, F. (2021, May). RpBERT: a text-image relation propagation-based BERT model for multimodal NER. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 15, pp. 13860-13868).

[7] Yu, J., Jiang, J., Yang, L., Xia, R. (2020). Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics.

[8] Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., Zhou, G. (2021, May). Multi-modal graph fusion for named entity recognition with targeted visual guidance. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 16, pp. 14347-14355).

[9] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... Krishnan, D. (2020). Supervised contrastive learning. Advances in neural information processing systems, 33, 18661-18673.

[10] Yang, X., Gong, X., Tang, B., Lei, Y., Deng, Y., Ouyang, H., ... Xu, Y. (2024, October). CAG: A Consistency-Adaptive Text-Image Alignment Generation for Joint Multimodal Entity-Relation Extraction. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 4183-4187).

[11] Li, J., Li, H., Pan, Z., Sun, D., Wang, J., Zhang, W., Pan, G. (2023). Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge. arXiv preprint arXiv:2305.12212.

[12] He, X., Li, S., Zhang, Y., Li, B., Xu, S., Zhou, Y. (2025). The more quality information the better: Hierarchical generation of multi-evidence alignment and fusion model for multimodal entity and relation extraction. Information Processing Management, 62(1), 103875.

[13] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[14] Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J. (2019). A fast and accurate one-stage approach to visual grounding. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4683-4693).

[15] Chen, Xiang, et al. "Hybrid transformer with multi-level fusion for multimodal knowledge graph completion." Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval. 2022.

[16] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[17] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

[18] Hsu, I. H., Huang, K. H., Boschee, E., Miller, S., Natarajan, P., Chang, K. W., Peng, N. (2022, July). DEGREE: A data-efficient generation-based event extraction model. In Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 1890-1908).

[19] Jia, M., Shen, L., Shen, X., Liao, L., Chen, M., He, X., ... Li, J. (2023, June). Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 7, pp. 8032-8040).

[20] OpenAI. GPT-3.5-Turbo [Large language model]. https://platform.openai.com/docs/models/gpt-3-5-turbo, 2022.

[21] Moon, S., Neves, L., Carvalho, V. (2018). Multimodal named entity recognition for short social media posts. arXiv preprint arXiv:1802.07862.

[22] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

[23] Zhang, Q., Fu, J., Liu, X., Huang, X. (2018, April). Adaptive co-attention network for named entity recognition in tweets. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[24] Zheng, C., Feng, J., Fu, Z., Cai, Y., Li, Q., Wang, T. (2021, October). Multimodal relation extraction with efficient graph alignment. In Proceedings of the 29th ACM international conference on multimedia (pp. 5298-5306).

[25] Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., Zhou, G. (2021, May). Multi-modal graph fusion for named entity recognition with targeted visual guidance. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 16, pp. 14347-14355).

[26] Liu, P., Wang, G., Li, H., Liu, J., Ren, Y., Zhu, H., Sun, L. (2024). Multi-granularity cross-modal representation learning for named entity recognition on social media. Information Processing Management, 61(1), 103546.

[27] Chen, Z., Li, Z., Liu, M., Zhang, C., Ma, H. (2025). Relevance-aware prompt-tuning method for multimodal social entity and relation extraction. Neurocomputing, 130316.

[28] Alibaba Cloud. Qwen-Max [Large language model]. https://www.aliyun.com/product/qwen, 2023.

[29] Meta Platforms, Inc. Meta-Llama-3.1-8B-Instruct [Large language model]. https://ai.meta.com/models/llama/, 2024.