CAT-MNER: MULTIMODAL NAMED ENTITY RECOGNITION WITH KNOWLEDGE-REFINED CROSS-MODAL ATTENTION

Xuwu Wang[*], *Jiabo Ye*[†], *Zhixu Li*[*,✉], *Junfeng Tian*[¶], *Yong Jiang*[¶], *Ming Yan*[¶], *Ji Zhang*[¶], *Yanghua Xiao*[*,‡,✉]

[*]School of Computer Science, Fudan University, China; [¶]Alibaba Group, China;
[†]School of Computer Science, East China Normal University, China;
[‡]Fudan-Aishu Cognitive Intelligence Joint Research Center, China.
{xwwang18,zhixuli,shawyh}@fudan.edu.cn, jiabo.ye@stu.ecnu.edu.cn
{tjf141457,yongjiang.jy,ym119608,zj122146}@alibaba-inc.com

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859972

ABSTRACT

Multimodal named entity recognition (MNER) aims to detect and classify named entities in multimodal scenarios. It requires bridging the gap between natural language and visual context, which presents two-fold challenges: the cross-modal alignment is diversified, and the cross-modal interaction is sometimes implicit. Existing MNER methods are vulnerable to some implicit interactions and are prone to overlook the involved significant features. To tackle this problem, we novelly propose to refine the cross-modal attention by identifying and highlighting some task-salient features. The saliency of each feature is measured according to its correlation with the expanded entity label words derived from external knowledge bases. We further propose an end-to-end Transformer-based MNER framework, which holds neater architecture yet achieves better performance than previous methods. Extensive experiments are conducted to validate the merits of our method. Moreover, our method reveals a significant advantage in data efficiency and generalization ability.

Index Terms— Multimodal Named Entity Recognition, Multimodal Interaction, Social Media

1. INTRODUCTION

Social media such as Twitter and Instagram provide massive user-generated content in the unstructured form such as texts and images. To extract relevant information from social media, multimodal named entity recognition (MNER) has attracted much attention, which aims at identifying and classifying[1] named entities from unstructured texts with attached images [**?**]. This task can be applied in many scenarios, such as multimedia relation extraction [1], multimedia search [2], etc.

[1]The classification types usually include person (PER), location (LOC), organization (ORG) and miscellaneous (MISC).

Fig. 1: MNER examples related with *Harry Potter*. We highlight corresponding named entities and present their types.

Compared with other domains such as newswire articles, texts on social media pose some inherent problems including colloquial language, short and coarse context, etc. So it is critical to incorporate the attached images to enhance the texts, which is however quite challenging. Besides the widely-concerned semantic gap problem, the obstacle usually lies in the complex cross-modal correlations between texts and images. *First, the cross-modal alignment is diversified.* As shown in Figure 1, "Harry Potter" is correlated with diversified multimodal contexts, including a person, a dog, a building, and film posters. This complex one-to-many correspondence thus brings great obstacles. *Second, the cross-modal interaction is sometimes implicit.* For example, the "Harry Potter" corresponds to the dog in Figure 1(b) based on the reasoning: *Harry Potter* $\xrightarrow{is\ a}$ *wizard* $\xleftarrow{dressed\ like}$ *dog*.

However, existing approaches are problematic for handling the complicated cross-modal interaction. Some efforts consider various mechanisms such as gated filter [3, 4] and cross-attention [5, 6] to model cross-modal interaction. Nevertheless, the cross-modal interaction is indirectly learned un-

der the supervision of MNER task annotations. So it is insufficient to learn the implicit cross-modal alignment such as the correspondence between the 'Harry Potter" and the "dog" in Figure 1(b). Some other works resort to image-text matching prediction to guide cross-modal interaction [7, 8], which only provides coarse-grained guidance and may even introduce noises. For instance, these models may predict the text and image in Figure 1(b) to be irrelevant. To summarize, without explicit and fine-grained cross-modal alignment annotations as the supervision signal, existing works can hardly capture the implicit cross-modal interaction. This leads to the neglect of some implicit features that are critical to comprehensively understand the multimodal context. And this problem becomes more pronounced when only limited training samples are provided.

To address this problem, we propose to resort to external knowledge to identify and highlight the task-salient features, and further refine the cross-modal attention. Intuitively, the saliency of a feature can be estimated according to its correlation with the entity labels, i.e., `PERSON`, `ORGANIZATION`, `LOCATION`, etc. However, these entity labels are far from enough to represent the comprehensive semantic information of entity categories. To tackle it, we propose to expand an entity label such as `PERSON` into a set of closely-relevant words such as "woman", "boy", "athlete", "hat", etc. by referring to either a textual knowledge base or a multimodal knowledge base such as WordNet and VisualGenome.

Based on the intuition above, we propose a novel framework named CAT-MNER, whose backbone is a Transformer with the refined cross-modal attention. Specifically, we first employ external knowledge to expand the entity type labels and use them to identify task-salient features. Then a gate mechanism utilizes the features' saliency scores to refine the cross-modal attention weights. This helps to highlight the task-salient features and meanwhile prevent them from being overlooked in the cross-modal interaction. Our main contributions are summarized as follows: 1) We newly propose to refine the cross-modal attention for MNER by identifying and highlighting task-salient features, where the saliency of each feature is measured with the help of external knowledge. 2) We further propose an end-to-end Transformer-based MNER framework, which holds neater architecture yet achieves better performance than previous methods. 3) Extensive experiments are conducted to validate the merits of our method. Specially, our method has the significant advantage of data efficiency and generalization ability.

2. RELATED WORK

There has been vast prior research about MNER on social media. Zhang et al. [3] first explored MNER in the tweets and proposed a co-attention network to incorporate the visual information. Since it is observed that *some visual signals are irrelevant with the texts and thus bring noises*, many works utilize the attention mechanism to extract text-related visual features while suppressing other visual information [4, 5, 9]. Besides, Sun et al. [7, 8] propagated image-text relation to the cross-modal fusion, Lu et al. [10] proposed to only resort to the multimodal NER model when the textual NER model is uncertain. To *address the semantic gap problem*, Zheng et al. [11] utilized adversarial learning to map features of different modalities to the same space, Wu et al. [12] used the object labels to represent visual features, Zhang et al. [6] designed a unified multimodal graph to represent the sentence and the image, Chen et al. [13] transformed images into captions.

3. METHODOLOGY

3.1. Overall Architecture

As shown in Figure 2(a), given a sentence T and an image I, the following three components are used for MNER.

Feature Extraction. For the sentence T, following the standard MNER approach, it is first tokenized and then mapped to a sequence of word embeddings $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_{L_T})$ with the embedding layer of a pre-trained language model (LM).

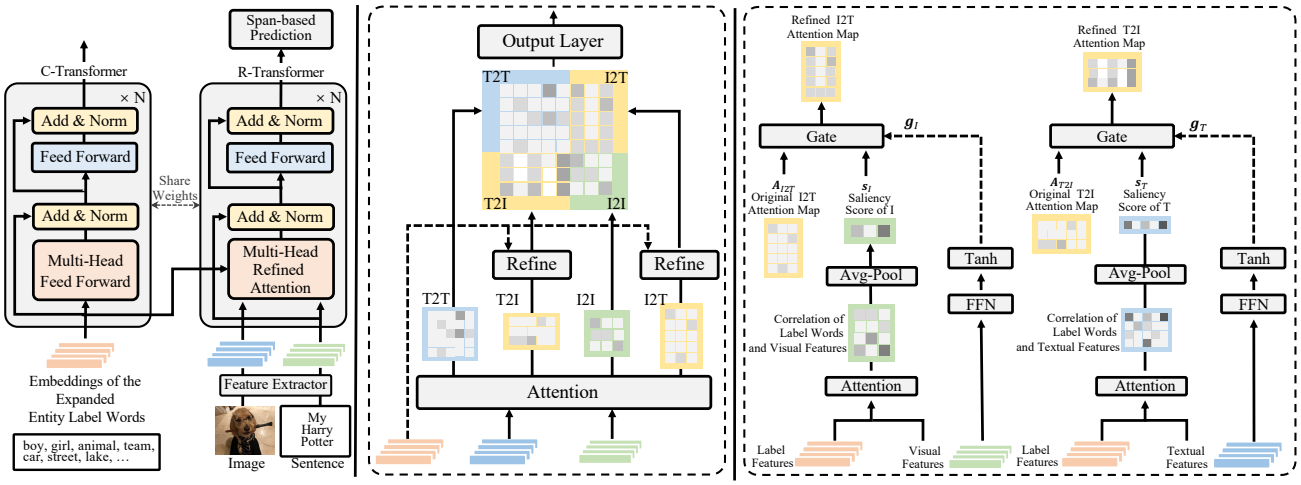
As for the image I, previous methods generally employ additional processing steps such as object detection. Unlike them, we utilize a pre-trained Vision Transformer (ViT) to extract visual features $(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_{L_I})$ in an end-to-end manner. Instead of directly feeding the image, we manually split the it into $L_I = 4 \times 4$ grids and then feed them into ViT.

Refined Cross-modal Attention. Here we need to obtain the word embeddings that are aware of the multimodal contexts. Conventional choices to model the cross-modal interactions include gated filter [3, 4], cross-attention [5, 6], etc., which might overlook some significant features involved in the implicit cross-modal relations. To prevent these features from being overlooked, we thus propose a R-Transformer$(\cdot)$ that modifies the cross-modal attention with some entity label words expanded from some knowledge bases (KBs).

$$\text{R-Transformer}\left(\{\boldsymbol{w}_i\}_{i=1}^{L_T}, \{\boldsymbol{v}_i\}_{i=1}^{L_I}, \{\boldsymbol{c}\}_{i=1}^{L_C}\right) \tag{1}$$

Here $\{\boldsymbol{c}\}_{i=1}^{L_C} = \text{C-Transformer}\left(\{c\}_{i=1}^{L_C}\right)$ are embeddings of the label words. Ideally, the R-Transformer$(\cdot)$ can stress the task-salient features, and further improve the cross-modal interactions than the vanilla cross-attention approach. The implementation details are elaborated in Section 3.2.

Span-based Prediction. Instead of the widely-applied CRF network, recent researches have proven the competitiveness of span-based NER models [14]. Following these works, we reformulate NER as the task of identifying the start and end indices of an entity span as well as assigning a category label to the span. So we first enumerate all possible spans in the sentence: $\{s_i\}_{i=1}^{N}$. Then for each text span $s_i = \{w_m, ..., w_n\}$, we concatenate the embeddings of the first and



(a) Overview (b) Multi-Head Refined Attention (c) Refine

Fig. 2: Framework of CAT-MNER

last tokens and then feed them into a feed-forward neural network (FFN) to predict its entity type: $l_i^c = \text{FFN}([\mathbf{w}_m||\mathbf{w}_n])$. Here l_i^c represents the logit of the entity type from {person, location, organization, misc, not_entity} [2].

Training Objective. Our objective is to assign a correct entity type to each enumerated span. So the loss function of MNER is formulated as the softmax cross-entropy loss:

$$p\left(l_i^c\right) = \frac{\exp(l_i^c)}{\sum_{\hat{c}=1}^{C} \exp(l_i^{\hat{c}})} \qquad \mathcal{L}_{\text{MNER}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \log p(l_i^c) \tag{2}$$

where C represents the number of the entity types. y_i^c is the binary ground-truth label of the i^{th} span.

3.2. Refined Multimodal Attention

Our R-Transformer$(\cdot)$ is the same as the general Transformer except for the refined multimodal attention component. So we elaborate its detailed design in this section. Considering the inputs come from both textual and visual modalities, namely T and I, we thus consider both intra-modal self-attention (T2T, I2I) and inter-modal cross-attention (T2I, I2T) as shown in Figure 2(b). To efficiently learn the complex inter-modal interactions, we propose to refine the cross-attention. In the following, we take the I2T attention, which aggregates information from visual features to each textual feature, as an example to illustrate our implementation. It is straight-forward to generalize to the T2I attention.

Overview. For a general I2T attention, textual and visual features are first projected into query (Q), key (K) and value (V) space. Then an attention mechanism is performed as below:

$$\boldsymbol{Q} = \text{FFN}_Q(\boldsymbol{w}),\ \boldsymbol{K} = \text{FFN}_K(\boldsymbol{v}),\ \boldsymbol{V} = \text{FFN}_V(\boldsymbol{v})$$
$$\text{Attention}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = \boldsymbol{A}_{I2T}\boldsymbol{V} = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V} \tag{3}$$

where $\boldsymbol{A}_{I2T} \in \mathbb{R}^{L_T \times L_I}$ represents the attention scores. $\boldsymbol{w} \in \mathbb{R}^{L_T \times d}$ and $\boldsymbol{v} \in \mathbb{R}^{L_I \times d}$ are embeddings of visual and textual tokens respectively. d is the shared dimension.

However, when the query and key belong to different modalities, the vanilla cross-attention may fail to capture some implicit correlations and assign a low attention weight to the corresponding interaction. Consequently, it is prone to overlook some important features. To highlight the task-helpful features and prevent them from being overlooked, we propose to utilize features' saliency scores to refine the attention scores as shown in Figure 2(c).

$$\boldsymbol{g_I} = \text{Tanh}\left(\text{FFN}\left(\boldsymbol{v}\right)\right)$$
$$\boldsymbol{A}'_{I2T} = \text{Softmax}\left(\frac{(1-\boldsymbol{g}_I)\odot\boldsymbol{Q}\boldsymbol{K}^\top + \boldsymbol{g}_I \odot \boldsymbol{s}_I}{\sqrt{d}}\right) \tag{4}$$

Here $\boldsymbol{Q}\boldsymbol{K}^\top$ is the original attention score. $\boldsymbol{s}_I \in \mathbb{R}^{L_I}$ represents the saliency score of visual tokens, which reveals which visual feature is more related to the MNER task. $\boldsymbol{g}_I \in \mathbb{R}^{L_I}$ is a gate used to control how much the saliency score should be considered. $\odot$ denotes the broadcasted element-wise multiplication. By combing the original attention score $\boldsymbol{Q}\boldsymbol{K}^\top$ and the saliency score $\boldsymbol{s}_I$ under the control of $\boldsymbol{g}_I$, the cross-modal interaction related with the salient features are enhanced.

Saliency Quantification. The critical problem now lies in obtaining each feature's saliency score, which is to quantify $\boldsymbol{s}_I$. To this end, we expand the entity type labels into specific words as the pivots: the correlation between the features and these expanded label words can measure their saliency.

[2]To avoid overlapping predictions, following [14], we sort the spans in the descending order of the logits and select non-overlap spans in order.

To achieve this purpose, inspired by [15], we expand each entity type label with the Top10 most related words. For the special label `MISC`, we first rewrite it with multiple specific entity labels, such as transportation, animal, etc., and then propose two methods to retrieve related words from a KB:

- **Retrieve from Textual KB.** We incorporate off-the-shelf Related Words to retrieve words having high semantic correlation with the entity type labels. Related Words is a KB that incorporates multiple KBs including WordNet, Concept Net, etc. For example, `person` is expanded with `worker, child, female`, etc.
- **Retrieve from Multimodal KB.** We also resort to the VisualGenome-based scene graph KB that records the multimodal knowledge of "object-predicate-object". Specifically, for each entity type label, we first select the Top5 related objects based on word vector similarity. Then we count the co-occurrence frequency of other objects directly connected with these seed objects in all scene graphs and select the Top5 frequent objects' labels as the related words. For example, `person` is expanded with `men, coat, arm`, etc.

Then to estimate the saliency score, given the expanded label word set $\{c_i\}_{i=1}^{L_C}$, we first embed them with a LM:

$$\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_{L_C}) = \mathcal{LM}(c_1, c_2, ..., c_{L_C}) \quad (5)$$

Considering the inter-modal attention is refined in each layer of the Transformer, it is necessary to keep the features of vision tokens and label words in the same feature space. Thus as shown in the left of Figure 2(a), we introduce a C-Transformer to encode the expanded label words. C-Transformer does not have the self-attention mechanism but shares other weights with the R-Transformer. For each layer in R-Transformer, the saliency score of visual tokens $\boldsymbol{s}_I \in \mathbb{R}^{L_I}$ is calculated through:

$$\boldsymbol{s}_I = \text{Average}\left(\text{Softmax}\left(\frac{\text{FFN}_Q(\boldsymbol{v})\,\text{FFN}_K(\boldsymbol{c})^\top}{\sqrt{d}}\right)\right) \quad (6)$$

where $\boldsymbol{c}$ comes from the C-Transformer, $\text{Average}(\cdot)$ denotes a mean pooling that reduces the second dimension.

Overall, estimating the saliency benefits from two types of knowledge: the symbolic knowledge from a KB and the continuous knowledge of pre-training tasks that helps to obtain the correlation score in Eq 6.

4. EXPERIMENT

4.1. Settings

Datasets. We test on two benchmarks: Twitter-2015 [3] and Twitter-2017 [4], which are also used by our predecessors.

Competitors. We compare with a wide range of baselines: 1) *For textual baselines*, we compare with BiLSTM-CRF [16], CNN-BiLSTM-CRF [17], BERT-CRF as well as the span-based NER models (e.g. BERT-span, RoBERTa-span) [14]. 2) *For multimodal baselines*, we compare with multiple state-of-the-art methods including OCSGA [12], UMT [5], IAIK [18], RpBERT [8] and UMGF [6].

Implementation Details. All the experiments are conducted on 8 NVIDIA V100 GPUs using Pytorch 1.7. To take advantage of the image-text matching ability obtained in pre-training, we use ViT from CLIP to extract visual features. We test with different pre-trained language modal backbones including RoBERTa and BERT. We set the AdamW optimizer with the learning rate of 5e-4 and use a warmup linear scheduler to control the learning rate. The batch size is set as 10.

4.2. Main Results

Multiple conclusions can be drawn based on the results in Table 1: 1) By comparing the textual baselines, we can see that the CRF-based methods and span-based methods achieve close results. And replacing the backbone of BERT with RoBERTa can further improve the performance. 2) Previous multimodal models do not consistently outperform the textual models [3]. We speculate that this is because inappropriate fusing the two modalities may even introduce noises and harm the prediction. 3) CAT-MNER achieves significant improvements compared with the existing methods. The improvement is more prominent on less frequent entity types (e.g., ORG, MISC in Twitter-2015 and MISC in Twitter-2017).

4.3. Detailed Analysis

Data Efficiency Analysis. To explore the model performance when training data is limited, we randomly sample $\alpha \in \{50, 100, 200, 400\}$ samples from the original training set (4000 and 3071 samples for Twitter-2015 and Twitter-2017) to train CAT-MNER and validate/test on the complete valid/test set. We report the results averaged over 8 samplings and runs for each α. For fairness, we also compare with the textual baseline (denoted as TEXT), CAT-MNER w/o the refinement mechanism (denoted as MM) and the UMT[4].

As shown in Figure 3, CAT-MNER significantly outperforms TEXT and UMT when the training data is extremely limited. But the improvements become smaller as the training data increases. Besides, compared with MM, CAT-MNER is more efficient for the less frequent entity types (e.g., ORG in Twitter-2015 and LOC, MISC in Twitter-2017), which reaffirms the superiority of our refinement mechanism.

Generalization Analysis. Table 2 shows the comparison of CAT-MNER and baselines for generalization analysis.

[3]The similar results were also observed by the concurrent work [19].

[4]It is the most advanced work that can be reproduced correctly.

Table 1: Performance comparison in Twitter-2015 and Twitter-2017. We report the Micro score of every evaluation metric. Results of methods with † come from [5]. Results with ‡ are reproduced by us. Other results are obtained from the corresponding papers. "e2e" indicates whether the model is end-to-end.

Modal.	Model	Twitter-2015							Twitter-2017							e2e
		F1 per type				Overall			F1 per type				Overall			
		PER	LOC	ORG	MISC	P	R	F1	PER	LOC	ORG	MISC	P	R	F1	
T	**BiLSTM-CRF**†	76.77	72.56	41.33	26.80	68.14	61.09	64.42	85.12	72.68	72.50	52.56	79.42	73.43	76.31	✔
	CNN-BiLSTM-CRF†	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37	✔
	BERT$_{base}$-CRF‡	85.37	81.82	63.26	44.13	75.56	73.88	74.71	90.66	84.89	83.71	66.86	86.10	83.85	84.96	✔
	BERT$_{base}$-SPAN‡	85.35	81.88	62.06	43.23	75.52	73.83	74.76	90.84	85.55	81.99	69.77	85.68	84.60	85.14	✔
	BERT$_{large}$-SPAN‡	85.37	82.58	66.06	46.22	76.29	75.56	75.92	92.83	86.76	84.50	71.52	87.31	87.05	87.18	✔
	RoBERTa$_{large}$-SPAN‡	87.20	83.58	66.33	50.66	77.48	77.43	77.45	94.27	86.23	87.22	74.94	88.71	89.44	89.06	✔
T+V	**OCSGA** [12]	84.68	79.95	56.64	39.47	74.71	71.12	72.92	-	-	-	-	-	-	-	✗
	UMT [5]	85.24	81.58	63.03	39.45	71.67	75.23	73.41	91.56	84.73	82.24	70.10	85.28	85.34	85.31	✔
	IAIK [18]	84.28	79.43	58.97	41.47	74.78	71.82	73.27	-	-	-	-	-	-	-	✗
	RpBERT [8]	-	-	-	-	-	-	74.40	-	-	-	-	-	-	-	✗
	UMGF [6]	84.26	83.17	62.45	42.42	74.49	75.21	74.85	91.92	85.22	83.13	69.83	86.54	84.50	85.51	✗
	CAT-MNER (BERT$_{base}$)	85.57	82.53	63.77	43.38	76.19	74.65	75.41	91.90	85.96	83.38	68.67	87.04	84.97	85.99	✔
	CAT-MNER (BERT$_{large}$)	86.28	83.44	**68.34**	46.85	77.16	76.61	76.89	92.59	86.43	86.52	73.65	88.70	87.12	87.90	✔
	CAT-MNER (RoBERTa$_{large}$)	**88.04**	**84.70**	68.04	**52.33**	**78.75**	**78.69**	**78.72**	**94.61**	**88.40**	**88.14**	**80.50**	**90.27**	**90.67**	**90.47**	✔

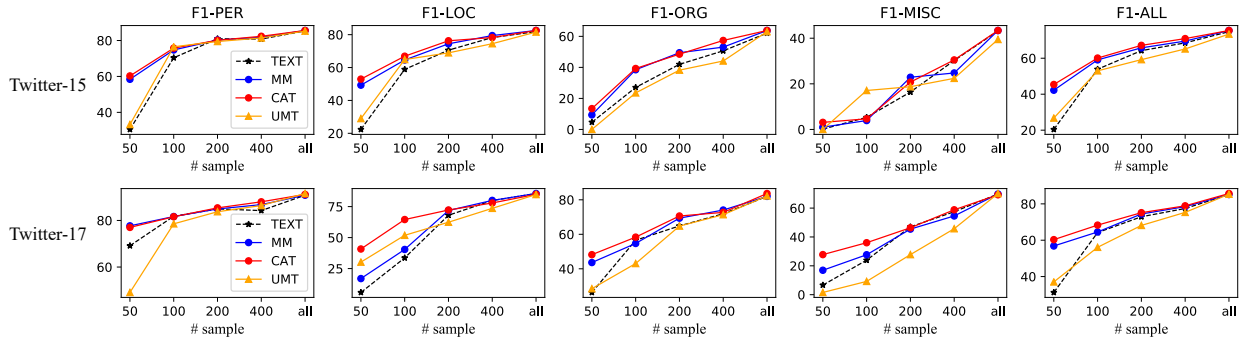


Fig. 3: Micro-F_1 of the methods under the data-limited condition. CAT represents CAT-MNER .

Table 2: Comparison of the generalization ability. Results with † come from [6].

Method	Twitter-17→Twitter-15			Twitter-15→Twitter-17		
	P	R	F1	P	R	F1
UMT†	64.67	63.59	64.13	67.80	55.23	60.87
UMGF†	67.00	62.81	66.21	69.88	56.92	62.74
CAT-MNER	**74.86**	**63.01**	**68.43**	**70.69**	**59.44**	**64.58**

Table 3: Ablation study. "w/o Refine." is implemented with self-attention over the concatenated multimodal inputs.

Method	Twitter-2015	Twitter-2017
Textual KB	78.72	90.47
w/o Refine.	77.63 (↓ 1.09)	89.23 (↓ 1.24)
w/o I2T Refine.	77.84 (↓ 0.88)	89.44 (↓ 1.03)
w/o T2I Refine.	78.08 (↓ 0.64)	89.77 (↓ 0.70)
w/o Image	77.45 (↓ 1.27)	89.06 (↓ 1.41)
Multimodal KB	78.22 (↓ 0.50)	90.13 (↓ 0.34)

All the models use BERT$_{base}$ as the text backbone. Twitter-17→Twitter-15 denotes the model is trained on Twitter-2017 and tested on Twitter-2015, and vice versa. We observe that CAT-MNER significantly outperforms the baselines by a large margin. This result demonstrates the strong generalization ability of our model.

Ablation Study. To verify the effectiveness of each component, we ablate them and report the results in Table 3. We can find out that: 1) For the refinement mechanism, it helps to improve the MNER performance a lot: with the textual baseline as the benchmark, CAT-MNER achieves a larger improvement than CAT-MNER w/o Refine. Besides, the I2T refinement is more important than the T2I refinement. The reason may be that MNER is a text-central task, so it is more critical to aggregate helpful visual information to enhance the textual information. 2) For different methods to expand the label words, it can be seen that textual KB-retrieved clues are

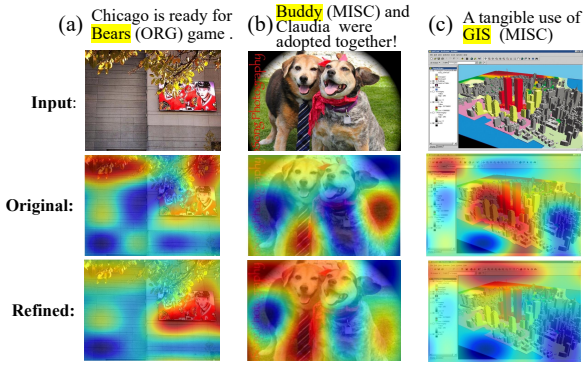


Fig. 4: Several cases about the attention map of the entity span highlighted in the text.

better than multimodal KB-retrieved ones. The reason may be that the latter have many fine-grained concepts words such as 'leg' and 'neck', which are not suitable for social media.

4.4. Case Study

We also present several examples of the attention map in Figure 4, which is used to explain the superiority of the refined attention for complex cross-modal interactions. 1) **For the one-to-many entity alignment**, although "bears" usually denote the animal that is correlated with the "tree" in Figure 4(a), the refined attention helps this textual span to attend to the visual feature of "athletes". Figure 4(b) shows similar results. 2) **For the unseen entities** such as the "GIS" in Figure 4(c), without refinement, it attends more to the building-like region in the picture. And the refinement mechanism successfully makes it attend to the menu of the application program.

5. CONCLUSION

In this paper, to address the problem of complicated cross-modal interactions in MNER, we propose to refine the attention scores with the features' saliency scores obtained from expanded entity label words. Both quantitative and qualitative experimental results verify the efficiency of our methods. Specially, we achieve huge performance boosts in terms of data efficiency and generalization ability.

Acknowledgement

This research was supported by the National Key Research and Development Project (No. 2020AAA0109302), National Natural Science Foundation of China (No. 62072323), Shanghai Science and Technology Innovation Action Plan (No. 19511120400), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0103), AECC Sichuan Gas Turbine Establishment (GJCZ-0033-19) and Alibaba Research Intern Program.

6. REFERENCES

[1] C. Zheng, Z. Wu, and et al., "Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts," in *ICME*, 2021.

[2] Z. Yang, W. Wang, H. Zhang, and B. Hu, "Learning homogeneous and heterogeneous co-occurrences for unsupervised cross-modal retrieval," in *ICME*, 2021.

[3] Q. Zhang and et al., "Adaptive co-attention network for named entity recognition in tweets," in *AAAI*, 2018.

[4] D. Lu and et al., "Visual attention model for name tagging in multimodal social media," in *ACL*, 2018.

[5] J. Yu, J. Jiang, and et al., "Improving multimodal named entity recognition via entity span detection with unified multimodal transformer," ACL, 2020.

[6] D. Zhang, S. Wei, and et al., "Multi-modal graph fusion for named entity recognition with targeted visual guidance," in *AAAI*, 2021.

[7] L. Sun, J. Wang, and et al., "Riva: A pre-trained tweet multimodal model based on text-image relation for multimodal ner," in *Coling*, 2020.

[8] L. Sun, J. Wang, and et al., "Rpbert: A text-image relation propagation-based bert model for multimodal ner," in *AAAI*, 2021.

[9] M. Seungwhan, N. Leonardo, and C. Vitor, "Multimodal named entity recognition for short social media posts," *NAACL-HLT*, 2018.

[10] L. Liu, M. Wang, and et al., "Uamner: uncertainty-aware multimodal named entity recognition in social media posts," *Applied Intelligence*, 2021.

[11] C. Zheng, Z. Wu, and et al, "Object-aware multimodal named entity recognition in social media posts with adversarial learning," *TMM*, 2020.

[12] Z. Wu, C. Zheng, and et al., "Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts," in *MM*, 2020.

[13] S. Chen, G. Aguilar, L. Neves, and T. Solorio, "Can images help recognize entities? a study of the role of images for multimodal ner," in *EMNLP*, 2021.

[14] I. Yamada, A. Asai, and et al., "Luke: deep contextualized entity representations with entity-aware self-attention," *EMNLP*, 2020.

[15] N. Ding, Y. Chen, and et al., "Prompt-learning for fine-grained entity typing," *arXiv:2108.10604*, 2021.

[16] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv:1508.01991*, 2015.

[17] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *ACL*, 2016.

[18] D. Chen, Z. Li, B. Gu, and Z. Chen, "Multimodal named entity recognition with image attributes and image knowledge," in *DASFAA*. Springer, 2021.

[19] X. Wang, M. Gui, and et al., "Ita: Image-text alignments for multi-modal named entity recognition," *arXiv:2112.06482v1*, 2021.