Multi-modal Graph Fusion for Named Entity Recognition with Targeted Visual Guidance

Dong Zhang,[1] Suzhong Wei,[2] Shoushan Li,[1*] Hanqian Wu,[2] Qiaoming Zhu,[1] Guodong Zhou[1]
[1] School of Computer Science and Technology, Soochow University, China
[2] School of Computer Science and Engineering, Southeast University, China
dzhang@suda.edu.cn, antcoder@outlook.com, lishoushan@suda.edu.cn, hanqian@seu.edu.cn,
{qmzhu, gdzhou}@suda.edu.cn

Abstract

Multi-modal named entity recognition (MNER) aims to discover named entities in free text and classify them into predefined types with images. However, dominant MNER models do not fully exploit fine-grained semantic correspondences between semantic units of different modalities, which have the potential to refine multi-modal representation learning. To deal with this issue, we propose a unified multi-modal graph fusion (UMGF) approach for MNER. Specifically, we first represent the input sentence and image using a unified multi-modal graph, which captures various semantic relationships between multi-modal semantic units (words and visual objects). Then, we stack multiple graph-based multi-modal fusion layers that iteratively perform semantic interactions to learn node representations. Finally, we achieve an attention-based multi-modal representation for each word and perform entity labeling with a CRF decoder. Experimentation on the two benchmark datasets demonstrates the superiority of our MNER model.

Introduction

Multi-modal named entity recognition (MNER) has become an important research direction in named entity recognition (NER) (Lu et al. 2018; Kruengkrai et al. 2020), due to its research significance in multi-modal deep learning and wide applications, such as structural extraction from massive multimedia news and web product information. It significantly extends the conventional text-based NER by taking images as additional inputs. The assumption behind this is that the structured extraction is expected to be more accurate than purely text-based NER, since the visual context helps resolve ambiguous multi-sense words (Zhang et al. 2020a; Ju et al. 2020).

Apparently, how to fully exploit visual information is one of the core issues in MNER, which directly impacts the model performance. To this end, a lot of efforts have been made, roughly consisting of: (1) encoding the whole image into a global feature vector (Figure 1(a)), which can be used to augment each word representation (Moon, Neves, and Carvalho 2018), or guide the word learning a vision-aware representation (Lu et al. 2018; Zhang et al. 2018) based on RNN framework; (2) segmenting the whole image averagely into multiple regions (Figure 1(b)) and make them interact with the text sequence based on Transformer framework (Yu et al. 2020).

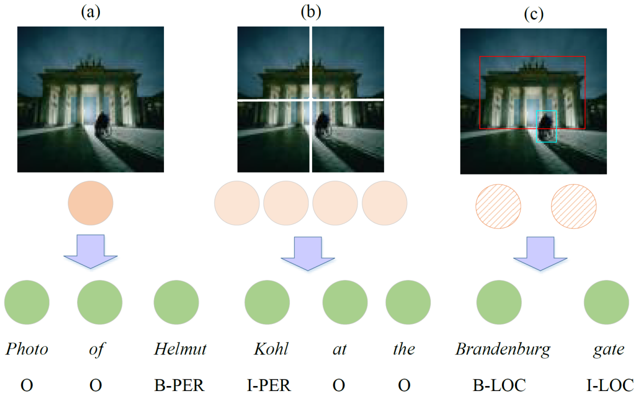


Figure 1: An example for multi-modal named entity recognition with (a) the whole visual cue, (b) averagely segmented visual cues and (c) targeted visual cues.

Despite their success, above studies do not fully exploit the fine-grained semantic correspondences between semantic units within an input sentence-image pair. For example, as shown in Figure 1, image (a) gives implicit global information, and image (b) includes local information of multiple averagely segmented regions, but it is still implicit. It is difficult for these two kinds of information to spread the clues of the ”gate” to the textual representations. The failed exploitation of this important clue may be due to two big challenges: 1) how to construct a unified representation to bridge the semantic gap between two different modalities, and 2) how to achieve semantic interactions based on the unified representation. However, image (c) gives the targeted object region as a clue. Relatively speaking, this kind of explicit information is more likely to help us identify some words as the correct entity type, such as “*Brandenburg gate*”. Therefore, we believe that such semantic correspondences can be exploited

[*] Corresponding Author

to refine multi-modal representation learning, since they enable the representations within one modality to incorporate cross-modal information as supplement during multi-modal semantic interactions.

To handle the above challenges, we propose a novel multi-modal graph fusion approach for MNER. We first represent the input sentence and image with a unified multi-modal graph. In this graph, each node indicates a semantic unit: textual word or visual object, and two types of edges are introduced to model semantic relationships between semantic units within the same modality (intra-modal edges) and semantic correspondences between semantic units of different modalities (inter-modal edges), respectively. Based on the graph, we then stack multiple graph-based multi-modal fusion layers that iteratively perform semantic interactions among the nodes to conduct graph encoding. Particularly, during this process, we distinguish the parameters of two modalities, and sequentially conduct intra- and inter-modal fusions to learn multi-modal node representations. Finally, we exploit these representations with a CRF decoder to perform entity labelling. Compared with previous models, ours is able to fully exploit semantic interactions among multi-modal semantic units for NER. Overall, the major contributions of our work are listed as follows:

- We propose a unified graph to represent the input sentence and image, where various semantic relationships between multi-modal semantic units can be captured for NER.

- We propose a unified multi-modal graph fusion approach (UMGF) to conduct graph encoding based on the above graph. To the best of our knowledge, our work is the first attempt to explore multi-modal graph neural network (GNN) for MNER.

- We conduct extensive experiments on Twitter 2015 and 2017 datasets, including both the self-domain and cross-domain investigation.

Multi-modal Graph Fusion for MNER

The encoder of our multi-modal NER model can be regarded as a multi-modal extension of GNN. We first represent the input sentence-image pair as a unified multi-modal graph. Then, based on this graph, we stack multiple multi-modal fusion layers to learn node representations, which provide the attention-based multi-modal context representation to the CRF decoder. In this section, we first introduce the task definition, then detail the specific components of our approach.

Task Definition. Given a sentence X and its associated image O as input, the goal of MNER is to extract a set of entities from X, and classify each extracted entity into one of the pre-defined types. As most existing work on MNER, we formulate the task as a sequence labeling problem. Let $X = (x_1, x_2, \cdots, x_n)$ denote a sequence of input words, $O = (o_1, o_2, \cdots, o_n)$ denote a set of input objects, and $y = (y_1, y_2, \cdots, y_n)$ be the corresponding label sequence, where $y_i \in Y$ and Y is the pre-defined label set with the BIO2 tagging schema (Sang and Veenstra 1999).

Multi-modal Graph

In this section, we take the sentence and the image shown in Figure 2 as an example, and describe how to use a multi-modal graph to represent them. Formally, our graph is undirected and can be formalized as $G = (V, E)$, which is constructed as follows:

Node Construction. In the node set V, each node represents either a textual word or a visual object. Specifically, we adopt the following strategies to construct these two kinds of nodes: (1) We include all words as separate textual nodes in order to fully exploit textual information. For example, in Figure 2, the multi-modal graph contains totally eight textual nodes, each of which corresponds to a word in the input sentence; (2) We employ the Stanford parser to identify all noun phrases in the input sentence, and then apply a visual grounding toolkit (Yang et al. 2019) to detect bounding boxes (visual objects) for each noun phrase. Since it is difficult to use some words to detect all related objects, we also introduce four general words of our pre-defined entity types (i.e., misc, person, location and organization) to encourage discovering more objects[1]. Subsequently, all detected visual objects are included as independent visual nodes. Let us revisit the example in Figure 2, where we can identify two noun phrases "*Photo of Helmut Kohl*" and "*the Brandenburg gate*" from the input sentence. But we can only leverage "gate" to detect an object because of no objects related "photo". Then, we can use the general word "person" to detect another object. Therefore, the two visual objects are included into the multi-modal graph.

Edge Construction. To capture various semantic relationships between multi-modal semantic units for NER, we consider two kinds of edges in the edge set E: (1) Any two nodes in the same modality are connected by an intra-modal edge; and (2) Each textual node representing any noun phrase and the corresponding visual node are connected by an inter-modal edge. Besides, each word and the visual node detected by the general words are connected with an inter-modal edge. Back to Figure 2, we can observe that all visual nodes are connected to each other, and all textual nodes are fully-connected. However, the objects detected by the original words are connected with the noun phrase, and the objects detected by the introduced general words are connected with all original words, both by inter-modal edges.

Before inputting the multi-modal graph into the stacked fusion layers, we introduce two multi-layer perceptrons (MLP) with ReLU activation function to project different features from two modalities onto the same space. Specifically, for each textual node v_{x_i}, we define its initial state $H^{(0)}_{x_i}$ as the sum of its word embedding from BERT and position encoding (Vaswani et al. 2017), then followed by textual MLP to form dimension d. To obtain the initial state $H^{(0)}_{o_j}$ of the visual node v_{o_j}, we leverage the visual features from ResNet (Ren et al. 2015), then followed by visual MLP to form dimension d.

[1] If no objects are detected, the object representations will be set to zero vectors and the model will degenerate to Transformer-based encoding.

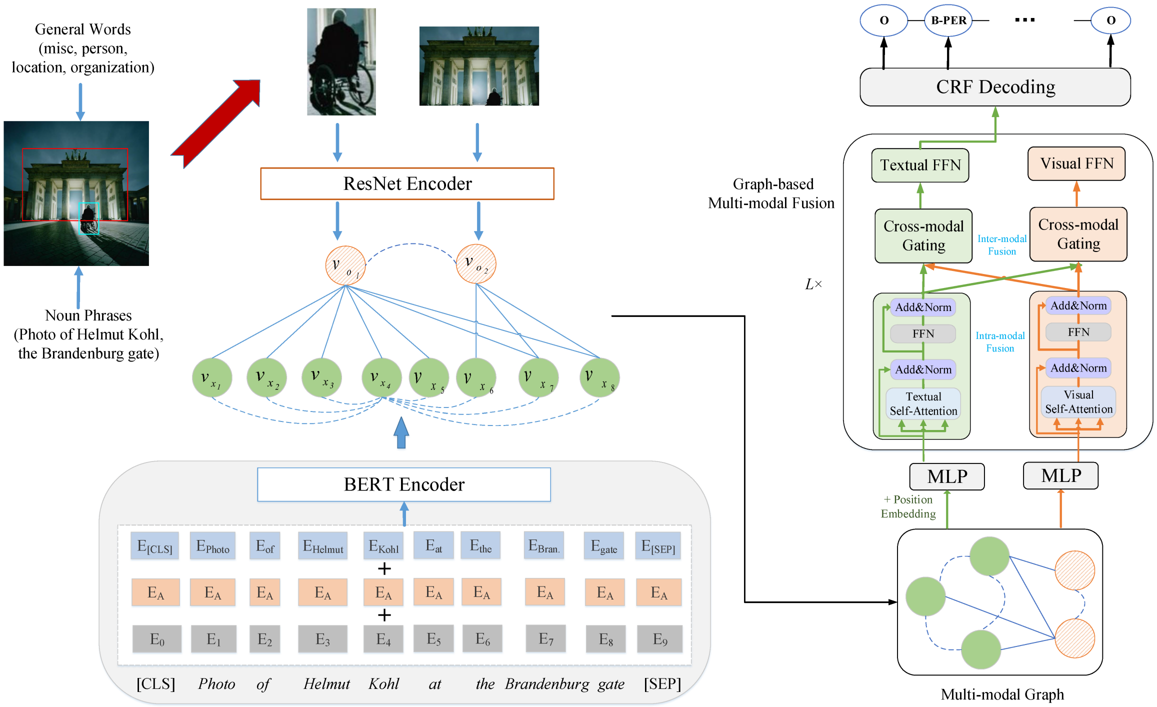


Figure 2: The overall architecture of our unified multi-modal graph fusion approach.

Graph-based Multi-modal Fusion

As shown in the right part of Figure 2, on the top of MLPs, we stack L graph-based multi-modal fusion layers to encode the above-mentioned multi-modal graph. At each fusion layer, we sequentially conduct intra- and inter-modal fusions to update all node states. In this way, the final node states encode both the context within the same modality and the cross-modal semantic information simultaneously. Particularly, since visual nodes and textual nodes are two types of semantic units containing the information of different modalities, we apply similar operations but with different parameters to model their state update process, respectively.

Specifically, in the l-th fusion layer, both updates of textual node states $\boldsymbol{H}_x^{(l)} = \{H_{x_i}^{(l)}\}$ and visual node states $\boldsymbol{H}_o^{(l)} = \{H_{o_j}^{(l)}\}$ mainly involve the following steps:

Intra-modal Fusion. We employ self-attention to generate the contextual representation of each node by collecting the message from its neighbors of the same modality. Formally, the contextual representations $\boldsymbol{C}_x^{(l)}$ of all textual nodes are calculated as follows:

$$\boldsymbol{C}_x^{(l)} = \mathrm{MultiHead}(\boldsymbol{H}_x^{(l-1)}, \boldsymbol{H}_x^{(l-1)}, \boldsymbol{H}_x^{(l-1)}) \qquad (1)$$

where $\mathrm{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ is a multi-head self-attention function taking a query matrix $\boldsymbol{Q}$, a key matrix $\boldsymbol{K}$, and a value matrix $\boldsymbol{V}$ as inputs. Similarly, we generate the contextual representations $\boldsymbol{C}_o^{(l)}$ of all visual nodes as:

$$\boldsymbol{C}_o^{(l)} = \mathrm{MultiHead}(\boldsymbol{H}_o^{(l-1)}, \boldsymbol{H}_o^{(l-1)}, \boldsymbol{H}_o^{(l-1)}) \qquad (2)$$

where we omit the descriptions of layer normalization and residual connection for simplicity.

Inter-modal Fusion. We apply a cross-modal gating mechanism with an element-wise operation to gather the semantic information of the cross-modal neighbours of each node.

Concretely, we generate the representation $R_{x_i}^{(l)}$ of a text node v_{x_i} in the following way:

$$R_{x_i}^{(l)} = C_{x_i}^{(l)} + \sum_{i \in \mathcal{N}(v_{x_i})} \alpha_{i,j} \odot C_{o_j}^{(l)} \qquad (3)$$

$$\alpha_{i,j} = \mathrm{sigmoid}(\mathrm{W}_1^{(l)} C_{x_i}^{(l)} + \mathrm{W}_2^{(l)} C_{o_j}^{(l)}) \qquad (4)$$

where $\mathcal{N}(v_{x_i})$ is the set of neighboring visual nodes of v_{x_i}, and $\mathrm{W}_1^{(l)}$ and $\mathrm{W}_2^{(l)}$ are parameter matrices.

Similarly, we produce the representation $R_{o_j}^{(l)}$ of a visual node v_{o_j} as follows:

$$R_{o_j}^{(l)} = C_{o_j}^{(l)} + \sum_{i \in \mathcal{N}(v_{o_j})} \beta_{j,i} \odot C_{x_i}^{(l)} \qquad (5)$$

$$\beta_{j,i} = \mathrm{sigmoid}(\mathrm{W}_3^{(l)} C_{o_j}^{(l)} + \mathrm{W}_4^{(l)} C_{x_i}^{(l)}) \qquad (6)$$

where $\mathcal{N}(v_{o_j})$ is the set of neighboring textual nodes of v_{o_j}, and $\mathrm{W}_3^{(l)}$ and $\mathrm{W}_4^{(l)}$ are also parameter matrices.

The advantage is that the above fusion approach can better determine the degree of inter-modal fusion according to the contextual representations of each modality. Finally, we adopt position-wise feed forward networks $\mathrm{FFN}(\cdot)$ to generate the textual node states $\boldsymbol{H}_x^{(l)}$ and visual node states $\boldsymbol{H}_o^{(l)}$:

$$\boldsymbol{H}_x^{(l)} = \mathrm{FFN}(\boldsymbol{R}_x^{(l)}) \qquad (7)$$

$$\boldsymbol{H}_o^{(l)} = \mathrm{FFN}(\boldsymbol{R}_o^{(l)}) \qquad (8)$$

Entity Type	Twitter-2015			Twitter-2017			Twitter-2017c		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Entities									
Person	2217	552	1816	2943	626	621	2586	626	621
Location	2091	522	1697	731	173	178	630	173	178
Organization	928	247	839	1674	375	395	1605	375	395
Misc.	940	225	726	701	150	157	629	150	157
Total	6176	1546	5078	6049	1324	1351	5450	1324	1351
Tweets									
Num.	4000	1000	3257	3373	723	723	3075	659	661

Table 1: The statistics summary of three Twitter datasets.

where $\boldsymbol{R}_x^{(l)} = \{R_{x_i}^{(l)}\}$ and $\boldsymbol{R}_o^{(l)} = \{R_{o_j}^{(l)}\}$ denote the above updated representations of all textual nodes and visual nodes respectively.

CRF Decoding

Since visual information has been incorporated into all textual nodes via multiple graph-based multi-modal fusion layers, we allow a decoder to perform conditional sequence labelling.

It has been shown that Conditional Random Fields (CRF) considers the correlations between labels in neighborhoods (Cao et al. 2018; Lison et al. 2020). Therefore, instead of decoding each label independently, we model them jointly using a CRF. Formally,

$$p(y|\bar{X}) = \frac{\prod_{i=1}^{n} \mathcal{S}_i(y_{i-1}, y_i, \bar{X})}{\sum_{y' \in Y} \prod_{i=1}^{n} \mathcal{S}_i(y'_{i-1}, y'_i, \bar{X})} \tag{9}$$

where $\mathcal{S}_i(y_{i-1}, y_i, \bar{X})$ and $\mathcal{S}_i(y'_{i-1}, y'_i, \bar{X})$ are potential functions. $\bar{X}$ is a generic input sequence with length of n.

In training phase, we use the maximum conditional likelihood estimation. Formally:

$$L(p(y|\bar{X})) = \sum_i \log p(y|\bar{X}) \tag{10}$$

Maximum conditional likelihood logarithm tries to learn parameters that maximize the log-likelihood $L(p(y|\bar{X}))$. In inference phase, we predict the output sequence that obtains the maximum score given by:

$$\hat{y} = \text{argmax}_{y' \in Y} p(y|\bar{X}) \tag{11}$$

Experimentation

We conduct experiments on two multi-modal NER datasets (Twitter 2015 and 2017) and a cleaned dataset from Twitter 2017, comparing our Unified Multi-modal Graph Fusion (UMGF) approach with a number of uni-modal and multi-modal approaches.

Experimental Settings

Datasets: Following (Yu et al. 2020), we first use two public Twitter datasets (i.e., Twitter-2015 and Twitter-2017) for MNER, which are provided by (Zhang et al. 2018) and (Lu et al. 2018), respectively. Since some samples in Twitter 2017 lack image modality[2], we remove these samples and obtain a cleaned version, namely Twitter-2017c. Table 1 shows the number of entities for each type and the size of data split.

Implementation Details. For each uni-modal and multi-modal approach compared in the experiments, the maximum length of the sentence input and the batch size are respectively set to 128 and 16. For our approach, the word embeddings X are initialized with the cased BERT_{base} model pre-trained by Devlin et al. (2019) with dimension of 768, and fine-tuned during training. The visual embeddings are initialized by ResNet152 with dimension of 2048 and fine-tuned during training. After MLPs, the dimension d of each node is transformed into 512. The head size in multi-head attention is set as 8. The learning rate, the dropout rate, and the tradeoff parameter are respectively set to 1e-4, 0.5, and 0.5, which can achieve the best performance on the development set of both datasets via a small grid search over the combinations of [1e-5, 1e-4], [0.1, 0.5], and [0.1, 0.9]. Based on best-performed development results, the layer number of multi-modal graph fusion is 2. To motivate future research, the code[3] will be released in our homepage.

Baselines

For a thorough comparison, we mainly compare two groups of baseline systems with our approach.

The first group are the representative text-based NER approaches: 1) **CNN-BiLSTM-CRF** (Ma and Hovy 2016), which is a classical text-based neural network for NER with both the word- and character-level information 2) **HBiLSTM-CRF** (Lample et al. 2016), which is an improvement of **CNN-BiLSTM-CRF**, replacing the bottom CNN layer with LSTM to build the hierarchical structure. 3) **BERT** (Devlin et al. 2019), which is most competitive baseline for NER with multi-layer bidirectional Transformer encoder and followed by stacking a softmax layer for entity prediction. 4) **BERT-CRF**, which is a variant of **BERT** by replacing the softmax layer with a CRF layer.

The second group are several competitive multi-modal approaches for MNER: 5) **VG** (Lu et al. 2018), which utilizes

[2] In the original Twitter-2017, the missing images of those samples are replaced with a uniform empty image.

[3] https://github.com/MANLP-suda/UMGF

Modality	Approaches	Twitter-2015							Twitter-2017						
		Single Type (F_1)				Overall			Single Type (F_1)				Overall		
		PER	LOC	ORG	MISC	$Pre.$	$Rec.$	F_1	PER	LOC	ORG	MISC	$Pre.$	$Rec.$	F_1
Text	CNN-BiLSTM-CRF	80.86	75.39	47.77	32.61	66.24	68.09	67.15	87.99	77.44	74.02	60.82	80.00	78.76	79.37
	HBiLSTM-CRF	82.34	76.83	51.59	32.52	70.32	68.05	69.17	87.91	78.57	76.67	59.32	82.69	78.16	80.37
	BERT	84.72	79.91	58.26	38.81	68.30	74.61	71.32	90.88	84.00	79.25	61.63	82.19	83.72	82.95
	BERT-CRF	84.74	80.51	60.27	37.29	69.22	74.59	71.81	90.25	83.05	81.13	62.21	83.32	83.57	83.44
Text +Image	VG	82.66	77.21	55.06	35.25	73.96	67.90	70.80	89.34	78.53	79.12	62.21	83.41	80.38	81.87
	ACoA	81.98	78.95	53.07	34.02	72.75	68.74	70.69	89.63	77.46	79.24	62.77	84.16	80.24	82.15
	UMT♣	**85.24**	81.58	**63.03**	39.45	71.67	**75.23**	73.41	91.56	84.73	82.24	**70.10**	85.28	**85.34**	85.31
	UMT	85.11	81.41	62.46	38.59	71.52	74.94	73.18	90.87	84.03	82.38	61.20	83.04	84.83	83.93
	UMGF (Ours)	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**†

Table 2: Performance comparison of different competitive uni-modal and multi-modal approaches for NER. For a fair comparison, we refer to the results of all baselines before UMT with the marker ♣ from (Yu et al. 2020). The marker † refers to significant test $p\text{-}value < 0.05$ when comparing with UMT.

Approaches	Twitter-2017c						
	Single Type (F_1)				Overall		
	PER	LOC	ORG	MISC	$Pre.$	$Rec.$	F_1
UMT	**90.83**	75.17	83.53	65.32	82.47	82.49	82.43
UMGF	89.67	**76.92**	**84.52**	**67.12**	**83.71**	**83.85**	**83.78**†

Table 3: Performance comparison of UMGF and UMT on Twitter 2017c dataset. The marker † refers to significant test $p\text{-}value < 0.05$ when comparing with UMT.

a visual attention and a gate mechanism to mine implicit information from a whole image to guide word representation learning based on **HBiLSTM-CRF**. 6) **ACoA**, which designs an adaptive co-attention network to learn word-aware visual representations and vision-aware word representations based on **CNN-BiLSTM-CRF**. (7) **UMT** (Yu et al. 2020), the state-of-the-art for MNER, which extends Transformer to multi-modal version and incorporates the auxiliary entity span detection module.

For all neural models, we conduct all the experiments on NVIDIA GTX 1080 Ti GPUs with pytorch 1.7.

Experimental Results

We mainly report the metric F_1 for each single type and overall precision ($Pre.$), recall ($Rec.$) and F_1 on two benchmark MNER datasets, as well as a cleaned dataset. To demonstrate the effectiveness and generalization of our approach, we conduct extensive experiments from self-domain and cross-domain scenarios.

Self-domain Scenario. Table 2 and 3 report the self-domain results. Specifically, Table 2 shows the performance comparison of different competitive uni-modal and multi-modal approaches for NER. From this table, we can see that:

1) For the uni-modal approaches, pre-trained approaches perform better than the conventional neural networks apparently. For example, **BERT-CRF** outperforms **HBiLSTM-CRF** by 2.4%, 3.68%, 8.68% and 4.77% on single type PER, LOC, ORG and MISC of Twitter-2015, respectively. This indicates obvious advantages of pre-training model in NER, which also explains that the recent approaches are typically based on BERT, such as **UMT**.

2) With regard to single type and overall results of both datasets, **BERT-CRF** with CRF decoding performs a little better than **BERT** except the metric $Rec.$. This shows the effectiveness of CRF layer for NER.

3) Compared with uni-modal approaches accordingly, multi-modal approaches achieve better performance generally. For example, in terms of overall F_1 on both datasets, **VG** outperforms **HBiLSTM-CRF** by 1.63% and 1.5%, respectively. Besides, **ACoA** outperforms **CNN-BiLSTM-CRF** by 3.54% and 2.78%, respectively. This suggests that incorporating visual context does facilitate NER task.

4) The most recent approach **UMT** performs much better than all uni-modal and other multi-modal baselines. We conjecture that the performance gains mainly come from the following reason: First, **UMT** leverage the Transformer structure to model textual sequence and perform cross-modal multi-head attention to learn more robust representation than conventional approaches. Second, this approach relies on an auxiliary entity span detection task, which provides an independent channel of text-based NER information.

5) Different from **UMT**, our proposed **UMGF** does not need to leverage an auxiliary task and build a unified graph to fully capture multi-modal semantic interactions between the textual and targeted visual nodes. Among all approaches, **UMGF** performs best and significantly outperforms **UMT** on both datasets.

Table 3 shows the performance comparison with the state-of-the-art **UMT** on a cleaned dataset Twitter-2017c. We can observe that **UMT** still performs much worse than **UMGF** except single type PER. This is due to the fact that our **UMGF** can fully utilize the enough visual guidance and maintain the stable performance, regardless of whether several images are missing.

Cross-domain Scenario. Table 4 shows performance comparison of UMT and UMGF in a cross-domain scenario for generalization analysis. For the first part, Twitter-2017 → Twitter-2015 denotes that the trained model on Twitter-2017 is used to test Twitter-2015. For the second part, Twitter-2015 → Twitter-2017 denotes that the trained model on Twitter-2015 is used to test Twitter-2017. From

Approaches	Twitter-2017 → Twitter-2015							Twitter-2015 → Twitter-2017						
	Single Type (F_1)				Overall			Single Type (F_1)				Overall		
	PER	LOC	ORG	MISC	$Pre.$	$Rec.$	F_1	PER	LOC	ORG	MISC	$Pre.$	$Rec.$	F_1
UMT	**80.34**	71.30	47.97	20.13	64.67	**63.59**	64.13	81.24	67.89	39.52	31.87	67.80	55.23	60.87
UMGF (Ours)	79.62	**71.94**	**49.48**	**20.24**	**67.00**	62.81	**66.21**†	**81.83**	**72.25**	**41.20**	**32.00**	**69.88**	**56.92**	**62.74**†

Table 4: Performance comparison of UMT and UMGF in a cross-domain scenario for generalization analysis. The marker † refers to significant test $p\text{-}value < 0.05$ when comparing with UMT.

Approaches	Twitter-2015							Twitter-2017						
	Single Type (F_1)				Overall			Single Type (F_1)				Overall		
	PER	LOC	ORG	MISC	$Pre.$	$Rec.$	F_1	PER	LOC	ORG	MISC	$Pre.$	$Rec.$	F_1
UMGF (Ours)	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**
UMGF w/o Tar.	82.56	79.89	58.70	36.96	71.12	71.30	71.21	90.02	82.05	79.27	67.12	84.34	82.20	83.26
UMGF w/o VCorr.	84.28	82.88	58.85	41.13	73.79	74.03	73.91	90.11	78.44	83.04	68.71	84.86	83.51	84.18
UMGF w/o Ind.	**84.65**	80.60	61.82	41.47	73.28	74.00	73.64	90.89	82.78	81.89	68.25	84.86	84.37	84.61

Table 5: Ablation study of our UMGF.

this table, we can see that our **UMGF** significantly outperforms **UMT** by a larger margin. For example, in terms of overall F_1, **UMGF** performs better than **UMT** by 2.08% and 1.87% on Twitter-2017 → Twitter-2015 and Twitter-2015 → Twitter-2017, respectively. The potential reasons that may influence the generalization are: (i) the biased labeling for different datasets; (ii) incomplete modalities of tweets in Twitter-2017; (iii) different type distribution of entities of two datasets.

Analysis and Discussion

Ablation Study. To investigate the importance of each component in our **UMGF**, we perform comparison between the full **UMGF** and its ablated approaches:

- **UMGF w/o Tar.**, a variant of our approach, which replaces the targeted visual guidance with average-segmented visual guidance.
- **UMGF w/o VCorr.**, a variant of our approach, which removes the correlations among visual objects as same as **UMT**.
- **UMGF w/o Ind.**, a variant of our approach, which adopts the shared parameters rather than independent for both modalities.

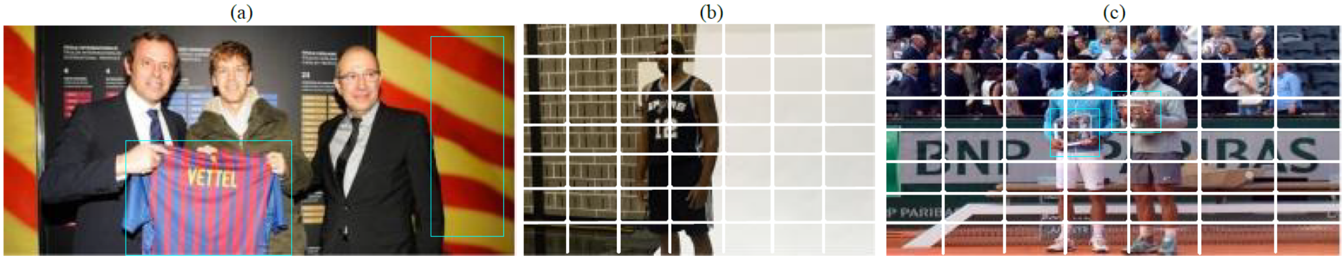
Table 5 shows the results of our full model **UMGF** and its ablated approaches. From this table, we can observe that although **UMGF** performs a little worse than **UMGF w/o Ind.** in terms of PER only, **UMGF** significantly outperforms two ablated approaches. Specifically, 1) compared with full model, **UMGF w/o Tar.** adopts the averagely segmented regions as nodes, rather than the targeted regions as nodes, which completely ignores the correspondences between semantic units of different modalities. Therefore, **UMGF w/o Tar.** brings in a significant performance degradation, which shows the importance of targeted visual guidance. 2) **UMGF w/o VCorr.** completely deletes the connections among visual nodes as same as **UMT**. In other words, this approach lacks the intra-modal interactions in visual modality. We can observe that without correlations among visual objects leads to performance loss, which indicates the usefulness of intra-modal dynamics among visual semantic units. 3) **UMGF w/o Ind.** assign different modalities with unified parameters so that it leads to an obvious performance drop, which indicates the validity of our approach using different parameters.

Case Study. From these cases, we want to convey several arguments, which support the effectiveness of our approach.

First, from Figure 3(a), we can see that although **BERT-CRF** correctly identifies the first entity, it gives a wrong prediction of *BARCELONA*. This is mainly because this approach completely neglects visual information, such as sport shirt and flag. On the contrary, **UMT** and **UMGF** can refer to partial visual regions so that they can accurately determine both the entities.

Second, we can observe from Figure 3(b) that according to textual modality only, **BERT-CRF** can correctly predict the types of both the entities due to its strong contextual learning. Surprisingly, **UMT** gives a wrong identification of the first entity *LaMarcus*, probably because it obtains the noise of accompanying image by segmenting the full region of a person, such as the wall and white curtain. Besides, the auxiliary task of entity span detection in **UMT** can only determine the boundary of each entity, but cannot help identify the type. While, our **UMGF** accurately classifies the types of both entities. This is due to the fact that **UMGF** can attend to the full visual region of person and guide the text to obtain a higher confidence, i.e., $0.476 > 0.335$.

Third, as shown in Figure 3(c), it is difficult to detect final three words as an accurate type from textual only because of separate context. For example, both **BERT-CRF** and **UMT** predict word *RG14* as a wrong label, i.e., O and ORG, respectively. On the one hand, **BERT-CRF** might learn the similar representations for words of PER, so give the accurate prediction of the final entities. On the other hand, **UMT** even misidentifies the third entity, more likely due to the influence of noisy visual regions. Therefore, the appearance of vague regions containing humans may not help the text



	(a)	(b)	(c)
	[*Vettel* PER]^1 *join to* [*BARCELONA* ORG]^2	*Your first look at* [*LaMarcus* PER]^1 *in a* [*Spurs* ORG]^2 *uniform* .	*Friendly rivalries are the best kind* . [*RG14* MISC]^1 [*Nadal* PER]^2 [*Djokovic* PER]^3
BERT-CRF	1-PER✓, 2-LOC×	1-PER✓(0.335), 2-ORG✓	1-O×, 2-PER✓, 3-PER ✓
UMT	1-PER✓, 2-ORG✓	1-MISC×, 2-ORG✓	1-ORG×, 2-PER✓, 3-O×
UMGF(Ours)	1-PER✓, 2-ORG✓	1-PER✓(0.476), 2-ORG✓	1-MISC✓, 2-PER✓, 3-PER✓

Figure 3: The first and second rows show several representative samples together with their manually labeled entities in the test set of our two Twitter datasets, and the bottom three rows show predicted entities of different approaches on these test samples.

to accurately identify the entity of PER, and it is important to exploit the image information correctly. Different from the above two approaches, in addition to the human regions, we also mark the regions containing certificate and trophy in the image. These two targeted visual clues help us correctly determine the first entity as MISC to a certain extent.

Related Work

As a crucial component of information extraction, named entity recognition (NER) has attracted much attention in the research community in the past two decades (Fu, Liu, and Zhang 2020; He et al. 2020; Liu et al. 2020). In the literature, early studies normally attempt to perform feature engineering and leverage different linear classifiers such as SVM, maximum entropy and CRF (Zhou and Su 2005; Zhang et al. 2006). In the past five years, deep learning approaches are demonstrated promising for NER (Li et al. 2020), such as CNN, LSTM and attention mechanism (Luo et al. 2015; Ma and Hovy 2016; Chiu and Nichols 2016; Yang, Liang, and Zhang 2018). Recently, pre-trained (Zhang et al. 2020b) and graph-based (Luo and Zhao 2020) approaches produce great improvement for sequence labelling tasks, especially NER. However, the above successful approaches are normally based on textual modality and formal text. While, the studies of NER with visual guidance (MNER) in social media are quite limited. Besides, there are no attempts for MNER with graph modeling. Therefore, in the following, we mainly overview relevant studies of MNER and the applications of graph neural networks (GNNs).

Multi-modal NER. To our best knowledge, Zhang et al. (2018), Moon, Neves, and Carvalho (2018) and Lu et al. (2018) are the first to explore this task in the same period. The main idea of their approaches is encoding the text through RNN and the whole image through CNN, then implicitly interacting the information of two modalities. Recently, Yu et al. (2020) leverage Transformer to model text sequences, and divide images equally for many vs.many cross-modal interaction. Besides, they introduce an auxiliary task of entity span detection to further improve performance.

Different from above studies, we represent the input sentence-image pair as a unified graph, where various semantic relationships between multi-modal semantic units can be effectively captured for multi-modal NER. Benefiting from the multi-modal graph, we further introduce an extended GNN to conduct graph encoding via multi-modal semantic interactions.

Graph Neural Networks. Recently, GNNs including gated graph neural network (Li et al. 2016), graph convolutional network (Kipf and Welling 2017) and graph attention network (Velickovic et al. 2018) have been shown effective in many tasks such as VQA (Li et al. 2019), emotion detection in conversations (Zhang et al. 2019), text generation (Song et al. 2019) and text representation (Xue et al. 2019). Close to our work, Yin et al. (2020) introduce GNN for multi-modal machine translation, but they rely on the aligned words and region by annotation, which are normally difficult to obtain.

Different from above studies, we utilize the noun phrases and general entity words to detect the targeted objects from the images of tweets. On this basis, each semantic unit of both modalities is represented as nodes. In this way, fine-grained multi-modal correspondences are achieved so as to refine the word representation for NER in a unified graph model. To the best of our knowledge, this is the first attempt to build a unified graph and novel stacked graph fusion approach for MNER.

Conclusion

In this paper, we propose a novel multi-modal graph fusion approach, which exploits various semantic relationships between multi-modal semantic units for MNER. Experiment results and analysis on both self-domain and cross-domain of two benchmark datasets demonstrate the effectiveness of our model.

In the future, we will explore more multi-modal tasks with targeted visual guidance via graph modeling, such as multi-modal opinion mining and multi-modal parsing in social media.

Acknowledgments

This work was supported by an NSFC grant No.62076176, and a Major Project of Ministry of Science and Technology of China No.2020AAA0108604. This work was also supported by a project funded by China Postdoctoral Science Foundation No.2020M681713, a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

We thank our anonymous reviewers for their valuable suggestions. We also appreciate Jiangsu Provincial Key Laboratory of Network and Information and Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education for offering the environment of our experiments.

References

Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; and Liu, S. 2018. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism. In *Proceedings of EMNLP 2018*, 182–192.

Chiu, J. P. C.; and Nichols, E. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguistics* 4: 357–370.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, 4171–4186.

Fu, J.; Liu, P.; and Zhang, Q. 2020. Rethinking Generalization of Neural Models: A Named Entity Recognition Case Study. In *Proceedings of AAAI 2020*, 7732–7739.

He, Q.; Wu, L.; Yin, Y.; and Cai, H. 2020. Knowledge-Graph Augmented Word Representations for Named Entity Recognition. In *Proceedings of AAAI 2020*, 7919–7926.

Ju, X.; Zhang, D.; Li, J.; and Zhou, G. 2020. Transformer-based Label Set Generation for Multi-modal Multi-label Emotion Detection. In *Proceedings of ACM MM 2020*, 512–520.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of ICLR 2017*.

Kruengkrai, C.; Nguyen, T. H.; Mahani, S. A.; and Bing, L. 2020. Improving Low-Resource Named Entity Recognition using Joint Sentence and Token Labeling. In *Proceedings of ACL 2020*, 5898–5905. URL https://www.aclweb.org/anthology/2020.acl-main.523/.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT 2016*, 260–270.

Li, H.; Xu, H.; Qian, L.; and Zhou, G. 2020. Multi-layer Joint Learning of Chinese Nested Named Entity Recognition Based on Self-attention Mechanism. In *Proceedings of NLPCC 2020*, 144–155.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Relation-Aware Graph Attention Network for Visual Question Answering. In *Proceedings of ICCV 2019*, 10312–10321.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. S. 2016. Gated Graph Sequence Neural Networks. In *Proceedings of ICLR 2016*.

Lison, P.; Barnes, J.; Hubin, A.; and Touileb, S. 2020. Named Entity Recognition without Labelled Data: A Weak Supervision Approach. In *Proceedings of ACL 2020*, 1518–1533.

Liu, S.; Sun, Y.; Li, B.; Wang, W.; and Zhao, X. 2020. HAMNER: Headword Amplified Multi-Span Distantly Supervised Method for Domain Specific Named Entity Recognition. In *Proceedings of AAAI 2020*, 8401–8408.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. In *Proceedings of ACL 2018*, 1990–1999.

Luo, G.; Huang, X.; Lin, C.; and Nie, Z. 2015. Joint Entity Recognition and Disambiguation. In *Proceedings of EMNLP 2015*, 879–888.

Luo, Y.; and Zhao, H. 2020. Bipartite Flat-Graph Network for Nested Named Entity Recognition. In *Proceedings of ACL 2020*, 6408–6418.

Ma, X.; and Hovy, E. H. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of ACL 2016*.

Moon, S.; Neves, L.; and Carvalho, V. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. In *Proceedings of NAACL-HLT 2018*, 852–860.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of NIPS 2015*, 91–99.

Sang, E. F. T. K.; and Veenstra, J. 1999. Representing Text Chunks. In *Proceedings of EACL 1999*, 173–179.

Song, L.; Gildea, D.; Zhang, Y.; Wang, Z.; and Su, J. 2019. Semantic Neural Machine Translation using AMR. *Trans. Assoc. Comput. Linguistics* 7: 19–31.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Proceedings of NIPS 2017*, 6000–6010.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *Proceedings of ICLR 2018*.

Xue, M.; Cai, W.; Su, J.; Song, L.; Ge, Y.; Liu, Y.; and Wang, B. 2019. Neural Collective Entity Linking Based on Recurrent Random Walk Network Learning. In *Proceedings of IJCAI 2019*, 5327–5333.

Yang, J.; Liang, S.; and Zhang, Y. 2018. Design Challenges and Misconceptions in Neural Sequence Labeling. In *Proceedings of COLING 2018*, 3879–3889.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A Fast and Accurate One-Stage Approach to Visual Grounding. In *Proceedings of ICCV 2019*, 4682–4692.

Yin, Y.; Meng, F.; Su, J.; Zhou, C.; Yang, Z.; Zhou, J.; and Luo, J. 2020. A Novel Graph-based Multi-modal Fusion

Encoder for Neural Machine Translation. In *Proceedings of ACL 2020*, 3025–3035.

Yu, J.; Jiang, J.; Yang, L.; and Xia, R. 2020. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In *Proceedings of ACL 2020*, 3342–3352.

Zhang, D.; Ju, X.; Li, J.; Li, S.; Zhu, Q.; and Zhou, G. 2020a. Multi-modal Multi-label Emotion Detection with Modality and Label Dependence. In *Proceedings of EMNLP 2020*, 3584–3593.

Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; and Zhou, G. 2019. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *Proceedings of IJCAI 2019*, 5415–5421.

Zhang, M.; Zhou, G.; Yang, L.; and Ji, D. 2006. Chinese Word Segmentation and Named Entity Recognition Based on a Context-Dependent Mutual Information Independence Model. In *Proceedings of ACL-SIGHAN 2006*, 154–157.

Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In *Proceedings of AAAI 2018*, 5674–5681.

Zhang, Z.; Wu, Y.; Zhao, H.; Li, Z.; Zhang, S.; Zhou, X.; and Zhou, X. 2020b. Semantics-Aware BERT for Language Understanding. In *Proceedings of AAAI 2020*, 9628–9635.

Zhou, G.; and Su, J. 2005. Machine learning-based named entity recognition via effective integration of various evidences. *Nat. Lang. Eng.* 11(2): 189–206.