Multimodal Named Entity Recognition via Co-attention-Based Method with Dynamic Visual Concept Expansion

Xiaoyu Zhao and Buzhou Tang(✉)

Harbin Institute of Technology (Shenzhen), Shenzhen, China

Abstract. Multimodal named entity recognition (MNER) that recognizes named entities in text with the help of images has become a popular topic in recent years. Previous studies on MNER only utilize visual features or detected concepts from a given image directly without considering implicit knowledge among visual concepts. Taking the concepts not detected but relevant to those in the image into consideration provides rich prior knowledge, which has been proved effective on other multimodal tasks. This paper proposes a novel method to effectively take full advantage of external implicit knowledge, called Co-attention-based model with Dynamic Visual Concept Expansion (CDVCE). In CDVCE, we adopt the concept co-occurrence matrix in a large-scale annotated image database as implicit knowledge among visual concepts and dynamically expand detected visual concepts conditioned on the concept co-occurrence matrix and the input text. Experiments conducted on two public MNER datasets prove the effectiveness of our proposed method, which outperforms other state-of-the-art methods in most cases.

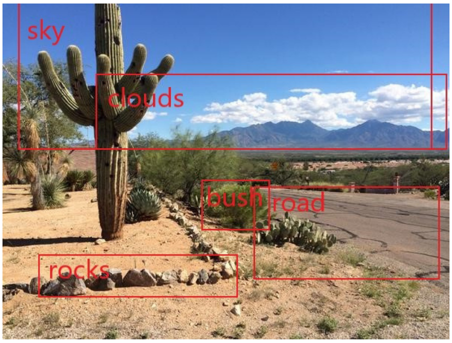
Keywords: Multimodal named entity recognition · External implicit knowledge · Co-occurrence matrix · Multimodal representation

1 Introduction

Named entity recognition plays a fundamental role in natural language processing field and serves as the cornerstone for many downstream tasks such as information extraction, question answering system, machine translating. Recent works argued that it is helpful to recognize textual named entities with the aid of visual clues [10,15,18]. With the emergence of social media platforms, a large amount of multimodal data generated by users has attracted much attention from researchers, and an increasing number of studies have been proposed for multimodal named entity recognition (MNER) in Twitter.

The current mainstream solution to MNER is to enhance the textual representation with visual information. Some researchers [9,10,14,20,22] employed image classification models such as ResNet [6] to extract region-level visual features as a supplement to text. However, there is an insuperable semantic gap

T. Mantoro et al. (Eds.): ICONIP 2021, LNCS 13108, pp. 476–487, 2021.
https://doi.org/10.1007/978-3-030-92185-9_39



@juliefowlis 1 mile from concert hall in **[LOC green valley]**.

Detected Concepts	Concepts with High Co-occurrence
sky	parasail, kite, moon, planes, tree branch, rainbow, clouds, sun, land, parachute, **valley**, mountain …
clouds	mountain, **valley**, sky, sun, moon, planes, range, horizon …
bush	tree branch, driveway, lawn, flower, yard, hedge, pathway, dirt road, grasses, twigs, shrubs, roses …
road	motorcycle, manhole, car, crosswalk, line, barricade, driveway, patch, town, tree branch …
rocks	town, mountain, **valley**, driveway, stream, polar bear, giraffes, rock wall, zebras, shore, enclosures …

Fig. 1. An example of MNER which needs the help of the implicit knowledge among visual concepts. On the left, there is a tweet with an associated image. On the right, there are detected visual concepts with the concepts frequently co-occur with them according to the visual concepts co-occurrence correlations, where the underlined concepts co-occur with more than one detected concept, the concepts in red appear in the image but are not detected and "LOC" indicates the entity type "Location".

between text and the region-level features of image. Some researchers [18,21] adopted object detection models to reduce the semantic gap between text and image. However, these approaches did not explore the implicit knowledge among visual concepts, which may limit the representation ability of image.

The implicit knowledge among visual concepts can be depicted by the co-occurrence relationships among visual concepts, the effectiveness of which has been proved on many other multimodal tasks, such as image-text matching [12,16], multi-label image recognition [2]. This idea is motivated by the fact that human beings are able to leverage outside-scene knowledge as supplements to comprehend images. For example, in Fig. 1, "sky", "clouds", "road", "bush" and "rocks" except "valley" are detected in the image due to the limitation of the object detection model [11], which could not provide enough visual information for MNER. According to our commonsense knowledge, "valley" co-occurs with "sky", "clouds" and "rocks" frequently. Therefore, we can employ the visual concept co-occurrence matrix to expand the visual concepts appropriately and make a correct inference. Introducing all co-occurred concepts of each detected concept may also introduce noises. For example, "planes" also appears frequently with "sky" and "clouds" but provides little information for MNER in Fig. 1 because they neither appears in the current scene and nor are related to the given text.

To take advantage of image effectively for MNER, we propose a novel method, called Co-attention-based method with Dynamic Visual Concept Expansion (CDVCE). CDVCE first adopts the concept co-occurrence matrix in a large-scale annotated image database as implicit knowledge among visual concepts and then dynamically expand detected visual concepts conditioned on the concept co-occurrence matrix and the input text. Given the detected concepts from a given image, we expand the concepts according to the co-occurrence matrix and regard the expanded ones as supplements. To avoid introducing too many noises from the expanded concepts, we design a Visual Concept Expansion algorithm to select concepts and a Text-guided Concept Self-Attention mechanism to extract the visual features conditioned on the associated words dynamically. The fusion of the expanded and original visual features are aggregated for each word through a multimodal co-attention mechanism.

In summary, our main contributions are listed as below.

- We make the first attempt to introduce external implicit knowledge in MNER task using the visual concepts co-occurrence matrix obtained from an additional large-scale annotated image database.
- We propose a novel Co-attention-based method with Dynamic Visual Concept Expansion (CDVCE) that effectively expands visual concepts.
- The experiments conducted on two public MNER datasets demonstrates that our method yields explainable predictions with better performance than other state-of-the-art methods in most cases.

2 Related Work

2.1 Multimodal NER

Recent years, lots of studies have shown that the performance of NER tasks can be improved by combining multimodal information such as textual and visual information. Early studies attempt to encode text through RNN-based methods and the image regions through CNN [9,10,22]. Moon et al. [10] proposed a multimodal attention module to selectively extract information from words, characters and images. Lu et al. [9] and Zhang et al. [22] incorporated the visual features into textual representation in the early stage and the late stage respectively. Despite the effectiveness of region-level visual information, Wu et al. [18] introduced object-level representation to extract entities precisely. With the great success of the pretrained model in natural language processing, Yu et al. [20] adopted BERT [3] as sentence encoder and ResNet [6] as image encoder, with a multimodal interaction module to capture alignments between words and image. Most previous works ignored the bias of unrelated image information, which causes the cross-modal attention mechanism to produce misleading cross-modality-aware representation. Yu et al. [20] leveraged text-based entity span detection as an auxiliary task to identify textual entities more precisely, which did not explore the relationship between text and image. Sun et al. [14,15] proposed a text-image relationship inference module which was based on a cross-modality transformer and trained on an labeled dataset to produce gated visual features. Zhang et al. [21] adopted text-guided object detection model to obtain text-related objects and introduced graph modeling in MNER tasks, which excluded those objects not relevant to text. However, the above methods only utilized the output of the visual feature extraction model while ignore the occluded or long tail concepts.

2.2 Multimodal Representation

A variety of multimodal fusion methods has been explored to generate multimodal representation.

Bilinear Fusion: To capture the correlations between words and image, bilinear pooling [5] produced multimodal features by computing their outer product, which resulted in generating n2-dimensional representation. To get higher efficiency, plenty of works were proposed and achieved better performance [1].

Cross-Attention-Based Methods: These approaches concatenated image and text as a sequence and fed it to the following feature extraction module. Such methods adopted a single-stream cross-modal transformer to learn deep interactions between two modalities and conducted pretraining tasks such as masked language modeling, masked region modeling, and image-text matching [13].

Co-attention-Based Methods: These approaches maintained sub-networks of all modalities to capture each single modality features and then aggregated features from other modalities [17]. Gao et al. [4] proposed a Dynamic Intra-modality Attention module to compute attention score within single modality conditioned on the information from the other modality. Zhang et al. [23] introduced semantics-based attention to capture latent intra-modal correlations.

3 Proposed Method

As most existing works in MNER, we formulate the problem as a sequence labeling task where given a sentence of n tokens $X = (w_1, w_2, \dots, w_n)$ with an image V, the goal of the task is to identify the correct spans of entities as well as their types in the sentence, represented by named entity labels $Y = (y_1, y_2, \dots, y_n)$.

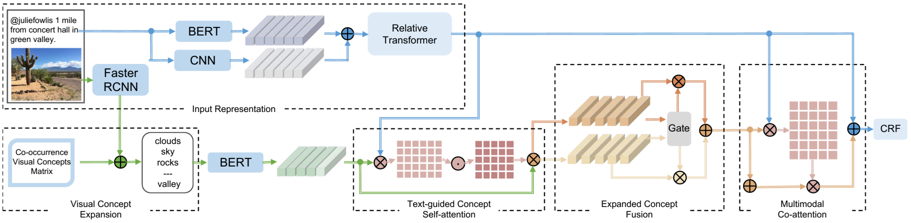


Fig. 2. Architecture of our proposed method CDVCE. It consists of six parts, Input Representation, Visual Concept Expansion, Text-guided Concept Self-attention, Expanded Concept Fusion, Multimodal Co-attention and CRF. The layer normalization and feed forward network are omitted for simplicity.

Figure 2 provides the detailed architecture of CDVCE for MNER. For sentence X with image V, we first use a BERT-based method to represent X and Faster R-CNN to detect objects in V respectively, and then expand the concepts about the detected objects dynamically via the Visual Concept Expansion, Text-guided Concept Self-Attention and Expanded Concept Fusion modules. Subsequently, the image-enhanced text representation is generated by the Multimodal Co-attention Module and fed to a CRF layer to make a prediction.

3.1 Input Representations

Textual Representation. We adopt pre-trained language model BERT[3] to encode the input sequence X, Meanwhile, to overcome the out-of-vocabulary (OOV) problem, CNN is adopted to capture the morphological information of each token. Next, we utilize relative transformer, a transformer-based encoder proposed by Yan et al. [19] with directional relative positional encoding and sharpened attention distribution, to obtain a direction- and distance-aware representation for each token. The representation of X is denoted as $\tilde{W} = [\tilde{w}_0, \tilde{w}_1, \cdots, \tilde{w}_n]$, where $\tilde{W} \in \mathbb{R}^{n \times d}$.

Visual Representation. In order to reduce the semantic gap between text and image, we first deploy Faster R-CNN [11] pretrained on Visual Genome [7], a fine-grained scene graph dataset with object, object attribute and object relationship annotations to detect objects from the input image V, and use the labels of the detected objects as the features of the image instead of the visual features generated by vision models in previous works for better interaction between words and visual concepts. Finally we utilize the embedding layer of BERT to encode these labels. Suppose that the objects detected from V by Faster R-CNN are denoted as $V = (v_0, v_1, \cdots, v_m)$ where m is the number of objects, and the visual representation of V is denoted as $\hat{V} = [\hat{v}_0, \hat{v}_1, \cdots, \hat{v}_m]$, where $\hat{V} \in \mathbb{R}^{m \times d}$.

3.2 Dynamic Visual Concept Expansion

Visual Concept Expansion. To explore knowledge among visual concepts, we expand the objects detected from image V according to the visual concept co-occurrence matrix based on relationships in Visual Genome. The co-occurrence matrix is constructed in the following way. When a relationship appears in an image, the corresponding objects are regarded as co-occurred concepts and the times of this object pair is increased by one. After scanning all images in Visual Genome, we regard the normalized times of all object pairs as the weights of all edges in the co-occurrence matrix. If two concepts appear frequently, the weight of the edge between them in the co-occurrence matrix is high, such as "sky" and "sun", "rocks" and "valley". We denote the visual concept co-occurrence matrix as C, where $C \in \mathbb{R}^{l \times l}$ and l is the size of visual concept vocabulary.

We design a simple rule to limit visual concept expansion to select concepts as close as possible to the detected concepts. Given the visual concepts detected by Faster R-CNN in image X, we retrieve their frequently co-occurrence concepts from C. For each detected concept vi , its frequently co-occurrence concept set is defined as $A_i = \{v_j | C_{ij} > t; v_i, v_j \in V_c, v_i \in V\}$ where t is the lower boundary of frequently co-occurrence weights and V_c is the concept vocabulary. If v_j appears in more than r frequently co-occurrence concept sets, we add it into the expanded visual concept set $V_e = (v_0, v_1, \cdots, v_m, v_{m+1}, v_{m+2}, \cdots, v_{m+e})$ where $(v_0, v_1, \cdots, v_m)$ is the original concepts and e is the number of expanded concepts. If we set r to 3, "planes", which is only related to "clouds" and "rocks" frequently, will not be added. The representation of the expanded visual concept set V_e is denoted as $\hat{V}_e = (\hat{v}_0, \hat{v}_1, \cdots, \hat{v}_{m+e})$ by BERT, where $\hat{V}_e \in \mathbb{R}^{(m+e)\times d}$.

Text-Guided Concept Self-attention. After the visual concept expansion, there is no doubt that we introduce potentially useful but undetected visual concepts into V_e as well as some useless concepts. To reduce helpless concepts, Text-guided Concept Self-attention is proposed to pay more attention on text-related visual concepts dynamically.

The naive self-attention mechanism only utilizes intra-modality information to estimate the object-to-object importance, ignoring the information from another modality. In MNER task, relations between the same visual concepts should have different weights conditioned on different associated text. So we modify the self-attention mechanism to pass the message from text to modeling the intra-modality interactions. First we calculate the similarities between visual concepts and textual tokens. The similarity matrices for V and V_e are:

$$\begin{gathered} S^{inter} = \tilde{W}^T K \hat{V}, \\ S_e^{inter} = \tilde{W}^T K_e \hat{V}_e, \end{gathered} \tag{1}$$

where $S^{inter}, S_e^{inter} \in \mathbb{R}^{m \times n}$ and $K, K_e \in \mathbb{R}^{d \times d}$ are the weighted matrices. Each element S_{ij}^{inter} in the matrix represents the similarity between visual concept v_i and textual token w_j and the i-th row of the matrix indicates the relevance vector between the whole sentence and concept v_i. The visual concept similarities to text stimulates the intra-modality attention Mechanism to focus more on the concepts with stronger semantic relation with the text and less on those text-unrelated but included in the expansion procedure. Following [23], we assume that if two visual concepts have similar response to the text, they are semantically related to each other. Therefore, we calculate the similarity between the concept-to-sentence relevance vectors to measure the similarity of the concepts themselves. The intra-modality similarity matrix S^{intra} can be calculated by

$$S^{intra} = S^{inter} K^{intra} S^{interT}, \tag{2}$$

where $K^{intra} \in \mathbb{R}^{\times}$ is the weighted matrix.

Then, we can update visual concept representation $\tilde{V}$, $\tilde{V}_e$ as follows:

$$\begin{gathered} \tilde{V} = softmax(S^{intra})\hat{V}, \\ \tilde{V}_e = softmax(S_e^{intra} \odot D_{V_e})\hat{V}_e. \end{gathered} \tag{3}$$

Finally, we apply the residual connection operation, layer normalization and feed forward network to construct a transformer-based block.

Expanded Concept Fusion. We regard the original detected visual concepts and the expanded ones as two independent channels. To fuse the complementary information within the different channels, following [8], we apply a gate mechanism to calculate the ratio to maintain the information from each channel as follows:

$$\begin{gathered} M = ReLU(FC([\tilde{V}, \tilde{V}_e, \tilde{V} - \tilde{V}_e, \tilde{V} \odot \tilde{V}_e])), \\ G = Sigmoid(FC(M)), \end{gathered} \tag{4}$$

where FC is fully-connected layer and $[\cdot, \cdot]$ denotes concatenation operation.

$$\mathring{V} = G \odot \tilde{V} + (1 - G) \odot \tilde{V}_e. \tag{5}$$

$\mathring{V}$ is the final output of Dynamic Visual Concept Expansion and G is the gate value.

3.3 Multimodal Co-attention

This module models the interactions between words and visual concepts to aggregate visual information for each word. We employ co-attention mechanism to generate image-enhanced textual representation $\mathring{W}$ as follows:

$$\begin{gathered} CoATT(Q, K) = softmax(\frac{QK^T}{\sqrt{d}})K, \\ \mathring{W}_{att} = LN(CoATT(\tilde{W}, \mathring{V}) + \tilde{W}), \\ \mathring{W} = LN(FFN(\mathring{W}_{att}) + \mathring{W}_{att}), \end{gathered} \tag{6}$$

where $LN(\cdot)$ denotes layer normalization and $FFN(\cdot)$ represents feed forward network. Finally, a Conditional Random Fields (CRF) layer is applied to improve tagging accuracy considering the correlations between neighbour labels by

$$\hat{y_i} = \arg\max_{y_i \in L} \frac{exp(E \cdot \mathring{W}_i + b)}{\sum_{y_{i-1} y_i} exp(E \cdot \mathring{W}_i + b)}, \tag{7}$$

where $\hat{y}_i$ is the prediction for i-th token, W and b are trainable parameters to model the transition from y_{i-1} to y_i.

4 Experiments

We conduct experiments on two public MNER datasets (Twitter 2015 [22] and Twitter 2017 [9]) to evaluate the effectiveness of our CDVCE method by comparing it with other state-of-the-art methods.

4.1 Datasets and Experimental Settings

Datasets. The tweets with their associated images in 2015 and 2017 respectively collected by Zhang et al. [22] and Lu et al. [9] through Twitter's API, that is, Twitter 2015 and Twitter 2017, are used for experiments. In both two datasets, four types of named entities including Person, Location, Organization and Misc are annotated. The statistics of the datasets are listed in Table 1.

Experimental Settings. In this study, we use the BIO scheme to represent named entities. The co-occurrence visual concept matrix generated from Visual Genome contains a concept vocabulary of 1,600 concepts. For textual representation, we use BERT-base-cased to initialize BERT. We set the max length of sentences as 64, the sizes of kernels in CNN as 2, 3 and 4, the number of layers of relative transformer as 2. For visual representation, the amount of concepts in one image detected by Faster R-CNN is set as 8. The lower boundary of frequently co-occurrence weights t is set as 0.01. Only the concepts in more than 3 frequently co-occurrence concept sets of the detected objects are selected for expansion. We set the number of expanded concepts as 32, the number of layers of the Text-guided Concept Self-attention Module as 2, the learning rate for BERT as 3e−5, the learning rate for the other modules as 1e−4, the dropout rate as 0.3, the weight decay as 0.05, and the batch size as 16 respectively. Precision (P), recall (R) and F1-score are used as performance evaluation metrics.

Table 1. Statistics of Twitter 2015 and Twitter 2017

Entity type	Twitter 2015			Twitter 2017		
	Train	Dev	Test	Train	Dev	Test
Person	2217	552	1816	2943	626	621
Location	2091	522	1697	731	173	178
Organization	928	247	839	1674	375	395
Misc	940	225	726	701	150	157
Total	6176	1546	5078	6049	1324	1351
Tweets	4000	1000	3257	3373	723	723

4.2 Results

We start with three baselines, i.e., **BiLstm-CRF**, **BERT-CRF** and **BERT-CRF-Rel**, that are state-of-the-art methods only using textual information, and then compare **CDVCE** with other state-of-the-art methods, i.e., **OCSGA**, **UMT**, **UMGF** and **RpBERT**, using both textual and visual information in different settings. A brief introduction to these methods is shown in as follows:

BiLstm-CRF is the method that only uses BiLstm to encode input text and CRF for label prediction. The word embeddings of words are initialized as Glove embeddings. **BERT-CRF** is the method that only uses BERT to encode input text and CRF for label sequence prediction. **BERT-CRF-Rel** is an extension of BERT-CRF that utilizes CNN to capture character features and adopt a 2-layer relative transformer [19] to obtain a direction- and distance-aware representation for each token. **OCSGA** [18] is the first multi-modal method to explore object-level features for MNER, which utilized BiLstm as text encoder and initialize object embeddings with the Glove embeddings.**UMT** [20] is the first multi-modal method to adopt transformer to encode text and image with a span detection task to alleviate the bias of incorporating visual features. **UMGF** [21] is another multi-modal method based on transformer that utilizes a text-guided object detection model to detect objects in the image and a unified graph to model the interactions between text and the targeted visual nodes. **RpBERT** is a multi-task method for MNER and text-image relation classification, where an external Twitter annotation dataset with text-image relations is used for text-image relation classification. RpBERT-base and RpBERT-large are two versions of RpBERT that use basic BERT and BERT-large pretrained on external large-scale Twitter data respectively.

Table 2. Comparison of different methods. The results of the methods marked with † are obtained from published papers, while that marked with ‡ are obtained by rerunning experiments using the official implementation.

Approaches	Mechanism	Twitter 2015			Twitter 2017		
		P	R	F1	P	R	F1
BiLstm-CRF	None	66.54	68.12	67.32	79.29	78.34	78.81
OCSGA† [18]	BiLstm-based + object-level	74.71	71.21	72.92	–	–	–
OCSGA‡ [18]	BiLstm-based + object-level	73.44	**71.71**	72.56	81.78	80.43	81.10
BiLstm-CRF + CDVCE	BiLstm-based + object-level	**74.93**	71.21	**73.02**	**83.52**	**80.68**	**82.07**
BERT-CRF	None	70.55	74.82	72.63	85.18	82.09	83.60
BERT-CRF-Rel	Relative transformer	71.59	74.90	73.21	83.95	85.20	84.57
UMT† [20]	Transformer-based + region-level	71.67	75.23	73.41	85.28	85.34	85.31
UMGF† [21]	Transformer-based + object-level	74.49	75.21	74.85	86.54	84.50	85.51
RpBERT-base† [15]	BiLstm + transformer-based + region-level	–	–	74.40	–	–	87.40
RpBERT-large† [15]	BiLstm + transformer-based + region-level	–	–	74.90	–	–	**87.80**
CDVCE (ours)	Transformer-based + object-level + knowledge	72.94	**77.66**	**75.23**	**86.57**	**87.79**	87.17

In order to illustrate the effectiveness of CDVCE, we apply the similar structure to BiLstm-CRF, that is, BiLSTM-CRF+CDVCE. The comparison of different methods on Twitter 2015 and Twitter 2017 is shown in Table 2.

It can be observed that the multi-modal methods are superior to uni-modal methods in MNER and CDVCE outperforms all other state-of-the-art methods on Twitter 2015 and all other state-of-the-art methods except RpBERT-large on Twitter 2017. The reason why CDVCE does not achieve better performance than RpBERT-large on Twitter 2017 is that the relation propagation based on the text-image relation classification task in RpBERT brings an significant improvement. As reported in [15], the relation propagation brings an improvement of 1.2 in F1 for RpBERT-base. Therefore, it is unfair to compare CDVCE with RpBERT that uses an external unpublic annotation dataset for another joint task. In spite of this, CDVCE outperforms RpBERT on Twitter 2015. Compared to the RpBERT method that used basic BERT the same as CDVCE, CDVCE achieves much higher F1 by about 0.8% on Twitter 2015. On Twitter 2017, CDVCE outperforms RpBERT-base without using relation propagation by about 0.9% (87.17% vs 86.2%) according to the results reported in [15]. Introducing relation propagation into CDVCE may also bring improvement, which is one direction of our future work.

Compared to OCSGA, BiLstm-CRF+CDVCE achieves higher F1 on both Twitter 2015 and Twitter 2017. Compared to UMGF, CDVCE achieces much higher F1 by 0.38% and 1.66% on Twitter 2015 and Twitter 2017 respectively. These results indicate that the proposed dynamic visual concept expansion is useful for MNER by introducing external knowledge and brings more improvement than the mechanisms for multi-modal information fusion in other methods.

4.3 Ablation Studies and Case Studies

We also conduct ablation studies on CDVCE by removing each module of dynamic visual concept expansion. Table 3 shows the results, where "w/o" denotes "without", "Expansion" denotes the visual concept expansion module, "Text-guidance" denotes the text-guided concept self-attention module, and "Fusion" denotes the expanded concept fusion module. Each module is beneficial

Table 3. Ablation study of our CDVCE.

Approaches	Twitter 2015			Twitter 2017		
	P	R	F1	P	R	F1
CDVCE(ours)	**72.94**	**77.66**	**75.23**	**86.57**	87.79	**87.17**
- w/o Expansion	73.16	76.22	74.66	85.43	87.64	86.52
- w/o Text-guidance	72.91	76.67	74.74	85.83	**87.86**	86.83
- w/o Fusion	72.37	77.36	74.78	84.67	87.49	86.06

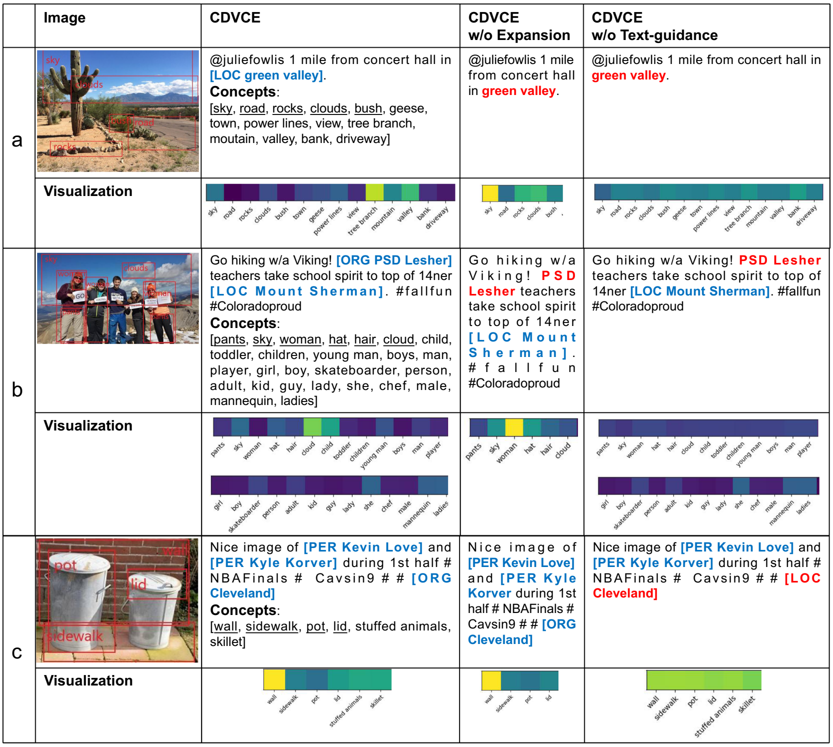


	Image	CDVCE	CDVCE w/o Expansion	CDVCE w/o Text-guidance
a		@juliefowlis 1 mile from concert hall in **[LOC green valley]**. **Concepts**: [sky, road, rocks, clouds, bush, geese, town, power lines, view, tree branch, moutain, valley, bank, driveway]	@juliefowlis 1 mile from concert hall in **green valley**.	@juliefowlis 1 mile from concert hall in **green valley**.
	Visualization			
b		Go hiking w/a Viking! **[ORG PSD Lesher]** teachers take school spirit to top of 14ner **[LOC Mount Sherman]**. #fallfun #Coloradoproud **Concepts**: [pants, sky, woman, hat, hair, cloud, child, toddler, children, young man, boys, man, player, girl, boy, skateboarder, person, adult, kid, guy, lady, she, chef, male, mannequin, ladies]	Go hiking w/a Viking! **PSD Lesher** teachers take school spirit to top of 14ner **[LOC Mount Sherman]**. #fallfun #Coloradoproud	Go hiking w/a Viking! **PSD Lesher** teachers take school spirit to top of 14ner **[LOC Mount Sherman]**. #fallfun #Coloradoproud
	Visualization			
c		Nice image of **[PER Kevin Love]** and **[PER Kyle Korver]** during 1st half # NBAFinals # Cavsin9 # # **[ORG Cleveland]** **Concepts**: [wall, sidewalk, pot, lid, stuffed animals, skillet]	Nice image of **[PER Kevin Love]** and **[PER Kyle Korver** during 1st half # NBAFinals # Cavsin9 # # **[ORG Cleveland]**	Nice image of **[PER Kevin Love]** and **[PER Kyle Korver]** during 1st half # NBAFinals # Cavsin9 # # **[LOC Cleveland]**
	Visualization			

sky
clouds
bush
road
rocks
sky
clouds
woman
GO
woman
sky road rocks clouds bush town geese power lines view tree branch mountain valley bank driveway
sky road rocks clouds bush .
sky road rocks clouds bush geese town power lines view tree branch mountain valley bank driveway
pants sky woman hat hair cloud child toddler children young man boys man player
girl boy skateboarder person adult kid guy lady she chef male mannequin ladies
pants sky woman hat hair cloud
pants sky woman hat hair cloud child toddler children young man boys man player
girl boy skateboarder person adult kid guy lady she chef male mannequin ladies
pot
lid
wall
sidewalk
wall sidewalk pot lid stuffed animals skillet
wall sidewalk pot lid
wall sidewalk pot lid stuffed animals skillet

Fig. 3. Examples tested by CDVCE in different settings. The detected concepts are highlighted in underline, the text fragments in blue are named entities correctly recognized, the text fragments in red are named entities not or wrongly recognized, and the concepts with visualized attentions in brighter colors are more relevant to the sentences. (Color figure online)

for CDVCE. Among the three modules of dynamic visual concept expansion, the visual concept expansion module is the biggest influencing factor on CDVCE. It proves that external knowledge is very import for MNER and CDVCE provides an effective way to take full advantage of external knowledge again. We further conduct the visualization analysis on several examples to prove the interpretability of our proposed method. The visual concepts in images detected by Faster RCNN, the named entities recognized by CDVCE in different settings, the concepts related to the detected visual concepts and the visualized attentions of the related concepts to sentences about three samples from the two datasets (a and b from Twitter 2015, and c from Twitter 2017) are shown in Fig. 3.

It is clear that the visual concept expansion module can expand related concepts effectively, and the text-guided concept self-attention module can model implicit relationships between images and sentences. For samples a and b, more entities are correctly detected by CDVCE because of the visual concept expansion module. For sample c, the location entity are correctly detected by CDVCE because of the Text-Guided Self-Attention module.

4.4 Conclusion

In this paper, we propose a multi-modal co-attention-based model with Dynamic Visual Concept Expansion (CDVCE) to incorporate implicit knowledge among images for the MNER task. The procedure of Dynamic visual Concept Expansion is implemented by three modules, that is, Visual Concept Expansion module, Text-guided Concept Self-Attention module and Expanded Concept Fusion module. Experiments on two public datasets proves the effectiveness of CDVCE with high interpretability.

Acknowledgements. This paper is supported in part by grants: National Natural Science Foundations of China (U1813215 and 61876052).

References

1. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: MUTAN: multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2612–2620 (2017)
2. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5177–5186 (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Gao, P., et al.: Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6639–6648 (2019)
5. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 317–326 (2016)

6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
8. Liu, L., Zhang, Z., Zhao, H., Zhou, X., Zhou, X.: Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. arXiv preprint arXiv:2009.06504 (2020)
9. Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H.: 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 (2018)
10. Moon, S., Neves, L., Carvalho, V.: Multimodal named entity recognition for short social media posts. arXiv preprint arXiv:1802.07862 (2018)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
12. Shi, B., Ji, L., Lu, P., Niu, Z., Duan, N.: Knowledge aware semantic concept expansion for image-text matching. In: IJCAI, vol. 1, p. 2 (2019)
13. Su, W., et al.: Vl-BERT: pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
14. Sun, L., et al.: RIVA: a pre-trained tweet multimodal model based on text-image relation for multimodal NER. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 1852–1862 (2020)
15. Sun, L., Wang, J., Zhang, K., Su, Y., Weng, F.: RpBERT: a text-image relation propagation-based BERT model for multimodal NER. arXiv preprint arXiv:2102.02967 (2021)
16. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12369, pp. 18–34. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58586-0_2
17. Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J.: Deep multimodal fusion by channel exchanging. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
18. Wu, Z., Zheng, C., Cai, Y., Chen, J., Leung, H.F., Li, Q.: Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1038–1046 (2020)
19. Yan, H., Deng, B., Li, X., Qiu, X.: TENER: adapting transformer encoder for named entity recognition. arXiv preprint arXiv:1911.04474 (2019)
20. Yu, J., Jiang, J., Yang, L., Xia, R.: Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. Association for Computational Linguistics (2020)
21. Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., Zhou, G.: Multi-modal graph fusion for named entity recognition with targeted visual guidance (2021)
22. Zhang, Q., Fu, J., Liu, X., Huang, X.: Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
23. Zhang, Q., Lei, Z., Zhang, Z., Li, S.Z.: Context-aware attention network for image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3536–3545 (2020)