Multimodal Named Entity Recognition with Image Attributes and Image Knowledge

Dawei Chen[1], Zhixu Li[1,2(✉)], Binbin Gu[4], and Zhigang Chen[3]

[1] School of Computer Science and Technology, Soochow University, Suzhou, China
dwchen@stu.suda.edu.cn, zhixuli@suda.edu.cn
[2] IFLYTEK Research, Suzhou, China
[3] State Key Laboratory of Cognitive Intelligence, iFLYTEK, Hefei, China
zgchen@iflytek.com
[4] University of California, Irvine, USA
binbing@uci.edu

Abstract. Multimodal named entity extraction is an emerging task which uses both textual and visual information to detect named entities and identify their entity types. The existing efforts are often flawed in two aspects. Firstly, they may easily ignore the natural prejudice of visual guidance brought by the image. Secondly, they do not further explore the knowledge contained in the image. In this paper, we novelly propose a novel neural network model which introduces both image attributes and image knowledge to help improve named entity extraction. While the image attributes are high-level abstract information of an image that could be labelled by a pre-trained model based on ImageNet, the image knowledge could be obtained from a general encyclopedia knowledge graph with multi-modal information such as DBPedia and Yago. Our emperical study conducted on real-world data collection demonstrates the effectiveness of our approach comparing with several state-of-the-art approaches.

Keywords: Named entity recognition · Multimodal learning · Social media · Knowledge graph

1 Introduction

Recent years have witnessed a dramatic growth of user-generated social media posts on various social media platforms such as Twitter, Facebook and Weibo. As an indispensable resource for many social media based tasks such as breaking news aggregation [20], the identification of cyber-attacks [21] or acquisition of user interests, there is a growing need to obtain structured information from social media. As a basic task of information extraction, Named Entity Recognition (NER) aims at discovering named entities in free text and classify them into

C. S. Jensen et al. (Eds.): DASFAA 2021, LNCS 12682, pp. 186–201, 2021.
https://doi.org/10.1007/978-3-030-73197-7_12

Tweet posts			
	teachers take on top of Mount Sherman.	Sony announced a Bad Boys in the next few years.	Jackson is really my favorite.
Expected NER results	teachers take on top of [Mount Sherman **LOC**].	[Sony **ORG**] announced a [Bad Boys **OTHER**] in the next few years.	[Jackson **PER**] is really my favorite.
NER with text only	teachers take on top of [Mount Sherman **OTHER**].	[Sony **ORG**] announced a [Bad Boys **PER**] in the next few years.	[Jackson **OTHER**] is really my favorite.
MNER with previous methods	teachers take on top of [Mount Sherman **PER**].	[Sony **PER**] announced a [Bad Boys **PER**] in the next few years.	[Jackson **OTHER**] is really my favorite.

Fig. 1. Three example social media posts with labelled named entities

per-defined types including person (*PER*), location (*LOC*), organization (*ORG*) and other (*OTHER*).

Different from NER with plain text, NER with social media posts is defined as multimodal named entity recognition (MNER) [29], which aims to detect named entities and identify their entity types given a (post text, post image) pair. As the three example posts with images given in Fig. 1, we expect to detect "Mount Sherman" as LOC from the first post, "Sony" as an ORG and "Bad Boys" as OTHER from the second post, and "Jackson" as PER from the third post. However, the post texts are usually too short to provide enough context for named entity recognition. As a result, if we perform named entity recognition with the post text only, these mentions might be wrongly recognized as shown in the figure. Fortunately, the post images may provide necessary complementary information to help named entity recognition.

So far, plenty of efforts have been made on NER. Ealier NER systems mainly rely on feature engineering and machine learning models [9], while the state-of-the-art approaches are sequence models, which replace the handcrafted features and machine learning models with various kinds of word embeddings [6] and Deep neural network (DNN) [3]. As a variant of NER, MNER also receives much attention in recent years [1,14,17,27,29]. Some work first learns the characteristics of each modality separately, and then integrates the characteristics of different modalities with attention mechanism [14,17,29]. Some other work produces interactions between modalities with attention mechanism in the early stage of extracting different modal features [1,27].

However, the existing MNER methods are often flawed in two aspects. Firstly, they may easily ignore the natural prejudice of visual guidance brought by the image. Let's see the first tweet post with an image in Fig. 1, where the "Mount Sherman" in the post might be taken as a person given that there are several

persons in the image. To allieviate the natural prejudice of visual guidance here, we need to treat the persons and the mountain in the image fairly, such that "Mount Sherman" is more likely to be associated with a mountain. Secondly, some important background knowledge about the image is yet to be obtained and furtherly explored. Let's see the second tweet post in Fig. 1, where the "Bad Boys" might be wrongly recognized as a person if we just use the shallow feature information in the image. But if we have the knowledge that the image is actually a movie poster, then "Bad Boys" in the text could be recognized as a movie instead of two persons. Similarly, we can see from the third post in Fig. 1, the director of the movie "Jackson" might be wrongly recognized as an aminal (i.e. OTHER), if we do not possess the knowledge that the image is a movie poster of the movie "King Kong".

To address the above drawbacks, we propose a novel MNER neural model integrating both *image attributes* and *image knowledge*. The image attributes are high-level abstract information of an image that are labelled by a pre-trained model based on ImageNet [22]. For instance, the labels to the image of the first post in Fig. 1 could be "person", "mountain", "sky", "cloud" and "jeans". By introducing image attributes, we could not only overcome the expression heterogeneity between text and image. More importantly, we could greatly alleviate the visual guidance bias brought by images. The knowledge about an image could be obtained from a general encyclopedia knowledge graph with multi-modal information (or MMKG for short) such as DBPedia [2] and Yago [24], which could be leveraged to better understand its meaning. However, it is nontrivial to obtain the image knowledge from MMKG, which requires us to find the entity that corresponds to the image in the MMKG firstly. It would be extreamly expensive if we search through the whole MMKG with millions of entities. Here we propose an efficient way to accomplish this task by searching the candidate entities corresponding to the entity mentions in the text, as well as their nearest neighbor entities within n-hop range.

To summarize, our main contributions are as follows:

- We introduce image attributes into MNER to alleviate the visual guidance bias brought by images and overcome the expression heterogeneity between text and image.
- We propose an efficient approach to obtain knowledge about a poster image from a large MMKG by utilizing the identified mentions in the poster text.
- We propose a novel neural model with multiple attentions to integrate both image attributes and image knowledge into our neural MNER model.

We conduct our empirical study on real-world data, which demonstrates the effctiveness of our approach comparing with several state-of-the-art approaches.

Roadmap. The rest of the paper is organized as follows: We discuss the related work in Sect. 2, and then present our approach in Sect. 3. After reporting our empirical study in Sect. 4, we finally conclude the paper in Sect. 5.

2 Related Work

In this section, we cover related work on traditional NER with text only, and MNER using image and text in recent years. Then, we present some other multi-modal tasks which inspire us deeply.

2.1 Traditional NER with Text only

The NER task has been studied for many years, and there are various mature models. Traditional approaches typically focus on designing effective features and then feed these features to different linear classifiers such as maximum entropy [5], conditional random fileds (CRF) [8] and support vector machines (SVM) [15]. Because traditional methods involve drab feature engineering, many deep learning methods for NER have emerged rapidly, such as BiLSTM-CRF [10], Bert-CRF [18], Lattice-LSTM [13]. It turns out that these neural approaches can achieve the state-of-the-art performance on formal text.

However, when using the above methods on social media tweets, the results are not satisfactory since the context of tweet texts is not rich enough. Hence, some studies propose to exploit external resources (e.g., shallow parser, Freebase dictionary and graphic characteristics) to help deal with NER task in social media text [11,12,33,34]. Indeed, the performance of these models with external resources is better than the previous work.

2.2 MNER with Image and Text

With the rapid increase of multi-modal data on social media platforms, some work starts to study using multi-modal data such as the associate images to improve the effectiveness of NER. Specifically, in order to fuse the textual and visual information, [17] proposes a multimodal NER (i.e. MNER) network with modality attention, while [29] and [14] propose an adaptive co-attention network and a gated visual attention mechanism to model the inter-modal interactions and filter out the noise in the visual context respectively. To fully capture intra-modal and cross-modal interactions, [1] extends multi-dimensional self-attention mechanism so that the proposed attention can guide visual attention module. Also, [27] proposes to leverage purely text-based entity span detection as an auxiliary module to alleviate the visual bias and designs a Unified Multimodal Transformer to guide the final predictions.

2.3 Other Multimodal Tasks

In the field of multimodal fusion, other multimodal tasks can also inspire us deeply. In VQA (Visual Question Answering) task, [25,31] introduces the attribute prediction layer as a method to incorporate high-level concepts. [4] proposes to introduce three modalities of image, text and image attributes for multi-modal irony recognition task in social media tweets. In [4], image attributes

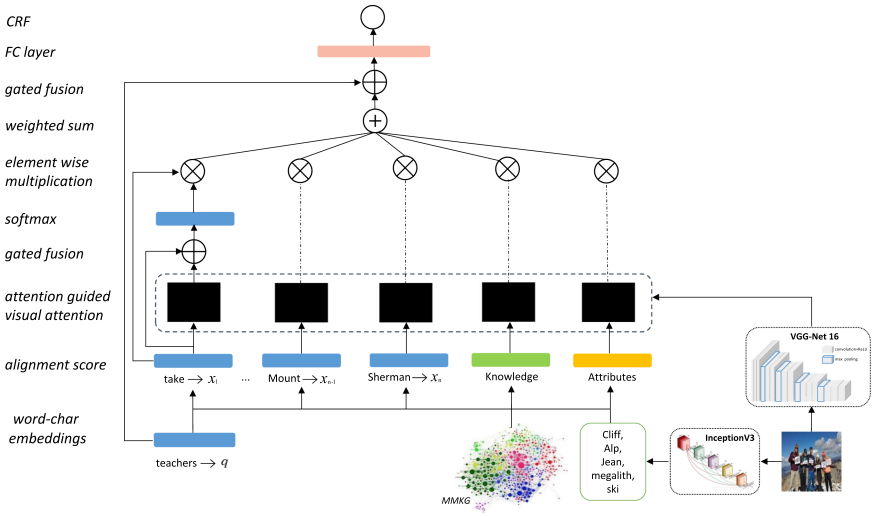


Fig. 2. The architecture of our proposed model

are used to ease the heterogeneity of image and text expression. The role of image attributes is a high-level abstract information bridging the gap between texts and images. [16,32] introduces knowledge to do some common sense reasoning and visual relation reasoning in Visual Question Answer task. Also, [23] proposes to combine external knowledge with question to solve problems where the answer is not in the image. While [7] designs modality fusion structure in order to discover the real importance of different modalities, several attention mechanisms are used to fuse text and audio [26,28]. An approach is proposed in [30] which constructs a domain-specific Multimodal Knowledge Graph (MMKG) with visual and textual information from Wikimedia Commons.

3 Our Proposed Model

In this work, we propose a novel neural network structure which includes the image attribute modality as well as image conceptual knowledge modality. This neural network uses an attention mechanism to perform the interaction among different modalities. The overall structure of our model is shown in Fig. 2. In the following, we first formulate the problem of MNER and then describe the proposed model in detail.

3.1 Problem Formulation

In this work, Multimodal Named Entity Recognition (MNER) task is formulated as a sequence labeling task. Given a text sequence $X = \{x_1, x_2, ..., x_n\}$ and associated image $Image$, MNER aims to identify entity boundaries from text

first with the BIO style, and also categorize the identified entities into predefined categories including Person, Organization, Location and Other. The output of a MNER model is a sequence of tags $Y = \{y_1, y_2, ..., y_n\}$ with the input text, where $y_i \in$ {O, B-PER, I–PER, B-ORG, I-ORG, B-LOC, I-LOC, B-OTHER, I-OTHER} in this work.

3.2 Introducting Image Attributes and Knowledge

Figure 2 illustrates the framework of our model. The model introduces image knowledge using Multimodal Knowledge Graph (MMKG) and image attributes using InceptionV3.[1] We describe each part of the model respectively next.

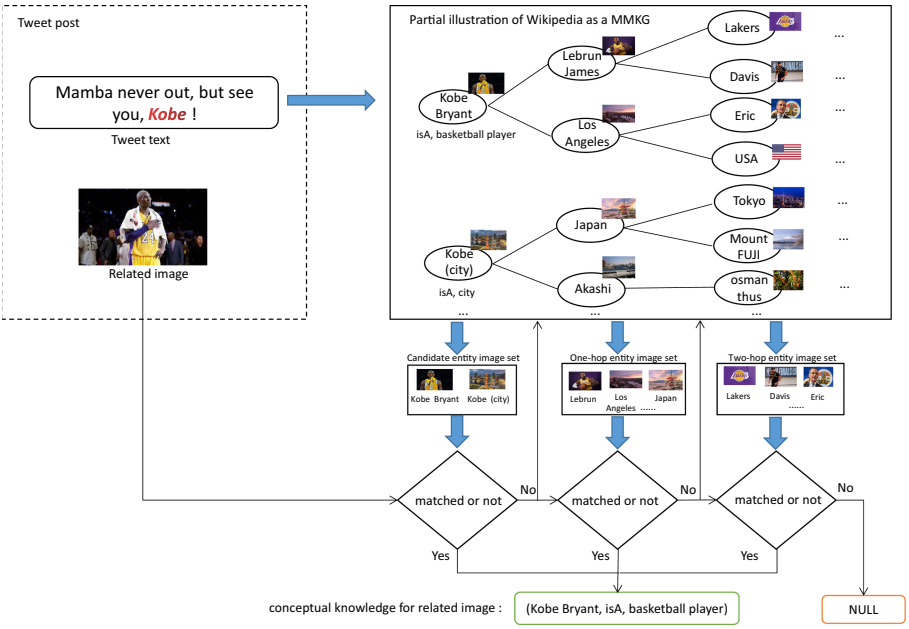


Fig. 3. The process of acquiring knowledge for an image from MMKG

Image Attributes. We use the InceptionV3 network pre-trained on ImageNet to predict the target objects in an image. Through the InceptionV3 network, we obtain the probability of a specific image corresponding to each category of 1000 categories in the ImageNet, and take the 5 category items with the highest probability value as the image attributes. Denote $IA(img)$ as the attribute set of the image img, we compute it as follows:

$$IA(img) = argsort\{p|p = InceptionV3(img)\}[1:5], p \in [0,1] \quad (1)$$

[1] Available at: https://keras.io/api/applications/#inceptionv3.

where *argsort* sorts 1000 probability values for the output of InceptionV3, and p is the probability score returned by InceptionV3.

Image Knowledge. As for obtaining the image knowledge, we use part of WikiPedia as the MMKG, which includes entities, the triple knowledge corresponding to the entity and the images corresponding to the entity. An example MMKG is given in the upper right part of Fig. 3.

To get knowledge for an image, a straightforward but very time-consuming way is to search through the entire MMKG to find the entity who owns the image that has the highest similarity to the given image. According to our observations, most of the time, the images are often closely related to the entities mentioned in the text. Thus, in this paper we propose an efficient way to acquire image knowledge by leveraging the (mention, candidate entity) pairs between the post text and MMKG. As shown in Fig. 3, from the input text, we first recognize entity mentions, i.e., "Kobe", and its corresponding candidate entities $S_e = \{e_1, e_2..., e_n\}$, i.e., "Kobe Bryant" and "Kobe (city)", from the MMKG according to the fuzzywuzzy algorithm.[2] Then we first calculate the similarity between the given image in the post and all the images of these candidate entities. If some highly similar image is found, the system would output the conceptual knowledge about its corresponding entity such as (Kobe Bryant, isA, basketball player). Otherwise, we get one-hop neighbourhood entities of the candidate entities, and find if any of these entities own similar images to the input image. If yes, we return relevant conceptual triplet as the image knowledge. Otherwise, we go to the two-hop neighbourhood entities of these candidate entities. But if no matched images are found even in the two-hop neighbourhood entities, we consider that these is probably no relevant knowledge about the image in the MMKG.

3.3 Feature Extraction

In this section, we use Convolutional Neural Network to extract character features and VGG network to extract image features.

Character Feature Extraction. Social media tweets are usually informal and contain many out-of-vocabulary (OOV) words. Character-level features could alleviate informal word and OOV problems because character features can capture valid word shape information such as prefixes, suffixes and capitalization. We use 2D Convolutional Neural Network to extract character feature vectors. First, a word w is projected to a sequence of characters $c = [c_1, c_2, ..., c_n]$ where n is the word length. Next, a convolutional operation of filter size $1 \times k$ is applied to the matrix $W \in \mathbb{R}^{d_e \times n}$. At the end, the character embedding of a word w is computed by the column-wise maximum operation.

Image Feature Extraction. In order to acquire features from an image, we use a pretrained VGG16 model. Specifically, we retain features of different image regions from the last pooling layer which has a shape of $7 \times 7 \times 512$ so that we can get the spatial features of an image. Moreover, we resize it to 49×512 to

[2] Available at: https://github.com/seatgeek/fuzzywuzzy.

simplify the calculations, where 49 is the number of image regions and 512 is the dimension of the feature vector for each image region.

3.4 Modality Fusion

In this section, we use attention and gated fusion module to combine text, attributes, knowledge and image information.

Self-attention. Self-attention module is applied to compute an alignment score between elements from the same source. In NLP (natural language processing), given a sequence of word embeddings $x = [x_1, x_2, ..., x_n]$ and a query embedding q, the alignment score $h(x_i, q)$ between x_i and q can be calculated using Eq. 2.

$$h(x_i, q) = w_t \sigma(x_i W_x + q W_q) \tag{2}$$

where σ is an activation function, w_t is a vector of weights and W_q, W_x are the weight matrices. Such an alignment score $h(x_i, q)$ evaluates how important x_i is to a query q. In order to refine the impact of each feature, we compute the feature-wise score vector $h'(x_i, q)$ in the following way.

$$h'(x_i, q) = W_t \sigma(x_i W_x + q W_q) \tag{3}$$

The difference between Eq. 2 and Eq. 3 is that $W_t \in \mathbb{R}^{d_e \times d_e}$ is a matrix and $h'(x_i, q) \in \mathbb{R}^{d_e}$ is a vector with the same length as x_i so that the interaction between each dimension of x_i and each dimension of q can be studied.

The purpose of softmax applied to the output function h' is to compute the categorical distribution $p(m|x, q)$ over all tokens. To reveal the importance of each feature k in a word embedding x_i, all the dimensions of $h'(x_i, q)$ need to be normalized and the categorical distribution is calculated as:

$$p(m_k = i|x, q) = softmax([h'(x_i, q)]_k) \tag{4}$$

where $[h'(x_i, q)]_k$ represents every dimension of $[h'(x_i, q)]$. Therefore, text context C for query q can be calculated as follows:

$$C = \left[\sum_{i=1}^{n} P_{ki} x_{ki} \right]_{k=1}^{d_e} \tag{5}$$

where $P_{ki} = p(m_k = i|x, q)$.

Alignment Score. Image attributes and image conceptual knowledge can be acquired by the approach described in Sect. 3.2. We concatenate image conceptual knowledge and image attributes to the end of the tweet text. For the tweet text and knowledge, the corresponding word vector representation can be directly obtained with fasttext. We use a two-layer fully connected network to obtain the vector representation of the image attribute embeddings based on the top 5 image attributes.

Denote a_s as the alignment score between a query embedding $q \in X$ and a word embedding w_i, we compute it as follows:

$$a_s = h'(w_i, q) \tag{6}$$

where $w_i \in X \cup K \cup A$ and $h'(w_i, q)$ can be calculated using Eq. 3 by substituting x_i with w_i. For the three sets X, K and A, $X = \{x_1, x_2, ..., x_n\}$ represents a set of word-char embedding of tweet text, K is the word-char embedding of knowledge and A means the word-char embedding of the weighted average of image attributes.

Attention Guided Visual Attention. To obtain the visual attention matrix, we calculate a_v between a_s and image feature matrix I as follows:

$$a_v(a_s, I_j) = W_v \sigma(a_s W_s + I_j W_i) \tag{7}$$

where $a_v(a_s, I_j) \in \mathbb{R}^{d_e}$ represents a single row of the visual attention scores matrix $a_v \in \mathbb{R}^{d_e \times N}$, $a_s \in \mathbb{R}^{d_e}$, $W_i \in \mathbb{R}^{d_i \times d_e}$, W_v, $W_s \in \mathbb{R}^{d_e \times d_e}$ are the weight matrices and $I_j \in \mathbb{R}^{d_i}$ is a row vector of $I \in \mathbb{R}^{d_i \times N}$.

Gated Fusion. We normalize the score a_v by Eq. 8 to get the probability distribution of a_v, denoted by $P(a_v)$, over all regions of image.

$$P(a_v) = softmax(a_v) \tag{8}$$

The output C_v containing visual context vector for a_s is an element-wise product between $p(a_v)$ and I which is computed as follows:

$$C_v = \sum_{i=1}^{n} P_i(a_v) \odot I_i \tag{9}$$

In order to dynamically merge alignment score a_s and visual attention vectors C_v, we choose a gate function G to integrate these information to get the fused representation F_r which is calculated as:

$$G = \sigma(W_1 a_s + W_2 C_v + b) \tag{10}$$

$$F_r = G \odot C_v + (1 - G) \odot a_s \tag{11}$$

where W_1 and W_2 are the learnable parameters and b is the bias vector and $\odot$ represents element-wise product operation.

We use Eq. 4 to get a categorical distribution P for F_r over all tokens of a sequence w, where w is a sequence of tweet text, knowledge and attributes. Then, element-wise product is computed between each pair of P_i and w_i for the purpose of getting context vector $C(q)$ for query q.

$$C(q) = \sum_{i=1}^{n} P_i \odot w_i \tag{12}$$

where n is the length of w, $C(q)$ is a context vector fused text, image, image attributes and knowledge features, $C(q) \in \mathbb{R}^{d_e}$.

To deal with textual attributes component of NER, we fuse word representation x with $C(q)$ with the gated fusion in Eq. 13 which is similar to Eq. 10. Later, we compute the final output O in the following way.

$$G = \sigma(W_1 C(q) + W_2 x + b) \tag{13}$$

$$O = G \odot C(q) + (1 - G) \odot x \tag{14}$$

3.5 Conditional Random Fields

Conditional Random Fields (CRF) is the last layer in our model. It has been shown that CRF is useful to sequence labeling task in practice because CRF can detect the correlation between labels and their neighborhood.

We take $X = \{x_0, x_1, ..., x_n\}$ as an input sequence and $y = \{y_0, y_1, ..., y_n\}$ as a generic sequence of labels for X. Y represents all possible label sequences for X. Given a sequence X, all the possible label sequences y can be calculated as follows:

$$p(y|X) = \frac{\prod_{i=1}^{n} \Omega_i(y_{i-1}, y_i, X)}{\sum_{y' \in Y} \prod_{i=1}^{n} \Omega_i(y'_{i-1}, y'_i, X)} \tag{15}$$

where $\Omega_i(y_{i-1}, y_i, X)$ and $\Omega_i(y'_{i-1}, y'_i, X)$ are potential functions. Maximum conditional likelihood logarithm is used to learn parameters to maximize the log-likelihood $L(p(y|X))$. The logarithm of likelihood is given by:

$$L(p(y|X)) = \sum_i log p(y|X) \tag{16}$$

At the time of decoding, we predict the output sequence y_o as the one with maximal score. The formula is shown as follows.

$$y_o = argmax_{y' \in Y} P(y|X) \tag{17}$$

4 Experiments

We conduct experiments on multimodal NER dataset and compare our model with existing unimodal and multimodal approaches. Precision, Recall and F1 score are used as the evaluation metrics in this work.

4.1 Dataset

We use multimodal NER dataset Twitter2015 constructed by [29]. It contains 4 types of entities Person, Location, Organization and Other collected from 8257 tweets. Table 1 shows the number of entities for each type in the train, validate and test sets.

Table 1. Details of dataset

	Train	Validate	Test
Person	2217	552	1816
Location	2091	522	1697
Organization	928	247	839
Other	940	225	726

4.2 Implementation Details

We use 300D fasttext[3] crawl embeddings to get the word embeddings. And we get 50D character embeddings trained from scratch using a single layer 2D CNN with a kernel size of 1×3. A pre-trained 16-layer VGG network is employed to initalize the vector representation of image. We set Adam optimizer with different learning rate: 0.001, 0.01, 0.03 and 0.005. The experimental results show that we achieve the best score when the learning rate is 0.001, the batch size is 20 and the dropout is 0.5. We adopt cosine similarity to compute the similarity among images and set threshold $\theta = 0.9$ to filter out dissimilar images.

4.3 Baselines

In this part, we describe four representative text-based models and multimodal models in comparison with our method.

- *BiLSTM-CRF*: BiLSTM-CRF was proposed by [8], requiring no feature engineering or data prepocessing. Therefore, it is suitable for many sequence labeling tasks. It was reported to have achieved great result on text-based dataset.
- *T-NER*: T-NER [19] is a specific NER system on tweet post. [29] applied T-NER to train a model on Twitter2015 training set and then evaluated it using Twitter2015 testing set.
- *Adaptive Co-Attention Network*: Adaptive Co-Attention Network was proposed by [29], which defined MNER problem and constructed the dataset Twitter2015.
- *Self-attention Network*: Self-attention Network was proposed by [1], which inspired us to use self-attention to capture the relationship among tweet text, image attributes and knowledge. This model achieved state-of-the-art effect on some metrics. Thus, we take this model as an important baseline to show the effectiveness of our model.

[3] Available at: https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip.

Table 2. Comparison of our approach with previous state-of-the-art methods.

	PER. F1	LOC. F1	ORG. F1	OTHER F1	Overall		
					Prec.	Recall	F1
BiLSTM+CRF [8]	76.77	72.56	41.33	26.80	68.14	61.09	64.42
T-NER [19]	83.64	76.18	50.26	34.56	69.54	68.65	69.09
Adaptive co-attention network [29]	81.98	78.95	53.07	34.02	72.75	68.74	70.69
Self-attention network [1]	83.98	78.65	**59.27**	39.54	73.50	**72.33**	72.91
Our model	**84.28**	**79.43**	58.97	**41.47**	**74.78**	71.82	**73.27**

4.4 Results and Discussion

In Table 2, we report the precision(P), recall(R) and F1 score(F1) achieved by each method on Twitter2015 dataset. In Table 3, we report the F1 score achieved by our method in two different scenarios: (1) With image and (2) Without image.

First, as illustrated in Table 2, by comparing all text-based approaches with multimodal approaches, it is obvious that multimodal models outperform the other models if the dataset only contains text. This indicates that visual context is indeed quite helpful for the NER task on social media tweet posts since image can provide effective information to enrich text context.

Second, as shown in Table 2, our method outperforms the baseline by 0.78% and 1.93% in LOC and OTHER types. Both overall precision and F1 of our method are better than that of the baselines. We assume that the improvement mainly comes from the following reason: the previous methods do not learn the real meaning of some images, whereas our approach can learn deep information of image and try to understand the really effective information that image can provide to text.

Third, although we have introduced the image attributes and knowledge, we still cannot remove the image from our model. From Table 3, we can see that if we remove image but introduce image attributes and knowledge, the F1 score of Without image is lower than that of With image scenario. This is because image attributes are unable to fully represent the image features and the information we need for some images is not deep conceptual knowledge but some target objects in images.

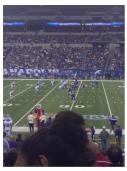
4.5 Bad Case Analysis

In Fig. 4, we show some examples where our approach fails for sequence labeling task. Some reasons are as follows:

Table 3. Results of our method with image and without image on our dataset.

	PER. F1	LOC. F1	ORG. F1	OTHER F1	Overall F1
With image	**84.28**	**79.43**	**58.97**	**41.47**	**73.27**
Without image	83.51	77.26	58.06	37.03	72.09

(a) [Reddit **ORG**] needs to stop pretending



(b) [Ben Davis **PER**] vs [Carmel **PER**]

Fig. 4. Two example wrong cases: (a) shows an unrelated image and a wrong prediction. (b) shows great ambiguity for text even if other information is introduced.

1) Unrelated image: Matched image do not relate with tweet text. As we can see in Fig. 4(a), "Reddit" belongs to "Other" but unrelated image could not provide valid information so that it results in wrong prediction "ORG".
2) Great ambiguity: Text is too short and has great ambiguity. As we can see in Fig. 4(b), "Ben Davis" and "Carnel" both belong to "ORG" but short tweet text has great ambiguity so that it is hard to help understand tweet even with some external information. Thus, it results in wrong prediction "PER".

5 Conclusions

In this paper, we propose a novel nerual network for multimodal NER. In our model, we use a new architecture to fuse image knowledge and image attributes. We propose an effective way to introduce image knowledge with MMKG to help us capture deep features of image to avoid error from shallow features. We introduce image attributes to help us treat the target objects in the image fairly alleviating the visual guidance bias of image naturally as well as expression heterogeneity between text and image. Experimental results show the superiority of our method compared to previous methods.

Future work includes two aspects. On the one hand, because our approach still performs not well on social media posts where text and image do not relate, we consider to identify the relevance of image and text and avoid introducing irrelevant image information to the model. On the other hand, since there are not many existing datasets and the size of the existing datasets is relatively small, we intend to build a larger and higher-quality dataset for this field.

Acknowledgment. This research is partially supported by National Key R&D Program of China (No. 2018AAA0101900), the Priority Academic Program Development of Jiangsu Higher Education Institutions, National Natural Science Foundation of China (Grant No. 62072323, 61632016), Natural Science Foundation of Jiangsu Province (No. BK20191420), and the Suda-Toycloud Data Intelligence Joint Laboratory.

References

1. Arshad, O., Gallo, I., Nawaz, S., Calefati, A.: Aiding intra-text representations with visual context for multimodal named entity recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 337–342. IEEE (2019)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Ives, Z.G.: DBpedia: a nucleus for a web of open data. In: Semantic Web, International Semantic Web Conference, Asian Semantic Web Conference, ISWC + ASWC, Busan, Korea, November (2007)
3. Bianco, S., Cadene, R., Celona, L., Napoletano, P.: Benchmark analysis of representative deep neural network architectures. IEEE Access **6**, 64270–64277 (2018)
4. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2506–2515 (2019)
5. Chieu, H.L., Ng, H.T.: Named entity recognition: a maximum entropy approach using global information. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002)
6. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(ARTICLE), 2493–2537 (2011)
7. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.: Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2018)
8. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. Computer Science. arXiv preprint arXiv:1508.01991 (2015)
9. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data (2001)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
11. Limsopatham, N., Collier, N.: Bidirectional LSTM for named entity recognition in twitter messages (2016)
12. Lin, B.Y., Xu, F.F., Luo, Z., Zhu, K.: Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media. In: Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 160–165 (2017)
13. Liu, C., Zhu, C., Zhu, W.: Chinese named entity recognition based on BERT with whole word masking. In: Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, pp. 311–316 (2020)
14. 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), vol. 1, pp. 1990–1999 (2018)

15. Luo, G., Huang, X., Lin, C.Y., Nie, Z.: Joint entity recognition and disambiguation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 879–888 (2015)
16. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: a visual question answering benchmark requiring external knowledge. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
17. Moon, S., Neves, L., Carvalho, V.: Multimodal named entity recognition for short social media posts. arXiv preprint arXiv:1802.07862 (2018)
18. Peng, M., Ma, R., Zhang, Q., Huang, X.: Simplify the usage of lexicon in Chinese NER. arXiv preprint arXiv:1908.05969 (2019)
19. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534 (2011)
20. Ritter, A., Etzioni, O., Clark, S.: Open domain event extraction from Twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1104–1112 (2012)
21. Ritter, A., Wright, E., Casey, W., Mitchell, T.: Weakly supervised extraction of computer security events from Twitter. In: Proceedings of the 24th International Conference on World Wide Web, pp. 896–905 (2015)
22. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
23. Su, Z., Zhu, C., Dong, Y., Cai, D., Chen, Y., Li, J.: Learning visual knowledge memory networks for visual question answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
24. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706 (2007)
25. Wu, Q., Shen, C., Liu, L., Dick, A., Van Den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 203–212 (2016)
26. Yang, Z., Zheng, B., Li, G., Zhao, X., Zhou, X., Jensen, C.S.: Adaptive top-k overlap set similarity joins. In: ICDE, pp. 1081–1092. IEEE (2020)
27. 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)
28. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X.: Hybrid attention based multimodal network for spoken language classification. In: Proceedings of the Conference. Association for Computational Linguistics. Meeting (2018)
29. Zhang, Q., Fu, J., Liu, X., Huang, X.: Adaptive co-attention network for named entity recognition in tweets. In: AAAI, pp. 5674–5681 (2018)
30. Zhang, X., Sun, X., Xie, C., Lun, B.: From vision to content: construction of domain-specific multi-modal knowledge graph. IEEE Access **7**, 108278–108294 (2019)
31. Zheng, B., et al.: Online trichromatic pickup and delivery scheduling in spatial crowdsourcing. In: ICDE, pp. 973–984. IEEE (2020)
32. Zheng, B., Su, H., Hua, W., Zheng, K., Zhou, X., Li, G.: Efficient clue-based route search on road networks. TKDE **29**(9), 1846–1859 (2017)

33. Zheng, B., Zhao, X., Weng, L., Hung, N.Q.V., Liu, H., Jensen, C.S.: PM-LSH: a fast and accurate LSH framework for high-dimensional approximate NN search. PVLDB **13**(5), 643–655 (2020)
34. Zheng, B., et al.: Answering why-not group spatial keyword queries. TKDE **32**(1), 26–39 (2020)