Neurocomputing 439 (2021) 12–21

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Hierarchical self-adaptation network for multimodal named entity recognition in social media

Yu Tian [a,b,c,d], Xian Sun [a,b,c,*], Hongfeng Yu [a,b], Ya Li [a,b], Kun Fu [a,b,c]

[a] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[b] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[c] University of Chinese Academy of Sciences, Beijing 100190, China
[d] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 21 May 2020
Revised 27 November 2020
Accepted 16 January 2021
Available online 21 January 2021
Communicated by Steven Hoi

2010 MSC:
00–01
99–00

Keywords:
Multimodal
Named entity recognition
Hierarchical self-adaptation network

ABSTRACT

Multimodal Named Entity Recognition task aims to identify named entities in user-generated posts containing both images and texts. Previous multimodal named entity recognition methods greatly benefit from visual features when the text and the image are well aligned, but this is not always the case in social media. On condition that the image is missing or mismatched with the text, these models usually fail to provide excellent performance. Besides, previous models use only single attention to capture the semantic interaction between different modalities, which largely ignore the existence of multiple entity objects in images and texts of the posts. To alleviate these issues, we present a novel model named Hierarchical Self-adaptation Network (HSN) to address these issues. The HSN contains 1) a Cross-modal Interaction Module to promote semantic interaction for the multiple entity objects in different modalities, which is proved to suppress wrong or incomplete attention in multimodal interactivity; 2) a Self-adaptive Multimodal Integration module to handle the problems that the images are missing or mismatched with the texts. Additionally, to evaluate the adaptability of HSN in real-life social media, we construct a Real-world NER dataset consisting of plain text posts and multimodal posts from Twitter. Extensive experiments demonstrate that our model achieves state-of-the-art results on the Real-world multimodal NER dataset and the Twitter multimodal NER dataset.

1. Introduction

Named Entity Recognition (NER) is a crucial task for Natural Language Processing (NLP), aims to identify named entities such as the name of person, place, and organization in the corpus. Since people combines texts, images, and videos for better individuality and expressiveness, social media posts become more multimodal and bring significant challenges to NER. As a result, Multimodal Named Entity Recognition (MNER) becomes the emerging task for NLP, which aims to identify named entities in multimodal posts and serves as an input role for other comprehensive tasks like Multimodal Machine Translation [1,2], Visual Dialog [3,4], and Multimodal Sentiment Analysis [5,6]. The textual part of the post usually provides limited contextual information [7], and the matching image can resolve this weakness by providing support information, which can be utilized to improve the performance of NER in multimodal scenarios. For example, as shown in Fig. 1, there are two multimodal posts about a named entity *Benz*. It is hard to tell which label that *Benz* belongs to via using textual information only. In contrast, using the matching image as a reference, we can easily distinguish that *Benz* is the name of the person in the first post and the car in the second post respectively.

As discussed above, we observe that semantic interaction between different modalities is the key of MNER. Although lots of works in MNER have been proposed in recent years with good performances [8–10], most of which have the same deficiencies can be further improved: 1) these works use only single attention to calculate the matching between multimodal features, which may lead to incomplete or wrong attention in multimodal interactivity, especially there are multiple entity objects existed in the textual feature or the image feature(sometimes in both); 2) these models via a single direction aided manner to fuse features and can not restrain the noise in the visual feature, therefore when the texts mismatch with the associated pictures, the visual features

* Corresponding author at: Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: sunxian@mail.ie.ac.cn (X. Sun).

https://doi.org/10.1016/j.neucom.2021.01.060
0925-2312/

Fig. 1. Example of how visual information can be helpful for NER. We can easily distinguish the label of Benz by the matching image.

will mislead the model and decrease the performance; 3) previous works are not suitable for real-life social media, when there is only the plain text existed in posts, as is the common scenario in social media, these works usually fail to provide satisfactory performance. This situation may be caused by that the blank image inputs act as interference in the processing of these works.

To handle these issues, we propose a novel model named Hierarchical Self-adaptation Network (HSN). Specifically, the model first extracts textual and visual features from posts via a Multimodal Feature Extractor. We then utilize a Cross-modal Interaction Module to learn semantic interaction between different modalities via Multi-head Hierarchical Attention (MHA). MHA can iteratively obtain relevant information between both features in different representations subspace to suppress incomplete or wrong attention in multimodal interactivity. Lastly, HSN uses a bi-directional manner called Self-adaptive Multimodal Integration to control the output weights of features from each module by itself. As a result, if the images are missing or mismatched with the texts, the textual feature will be the main contribution. Conversely, the fusion feature will be the main contribution. Nonetheless, all these schemes are designed and expected to against the mismatching issue in MNER tasks.

As discussed, there are both plain text posts and multimodal posts in social media, and an outstanding model should be able to process both accurately to apply to real-life social media. To evaluate the adaptability in the real-life scenario, i.e., the anti-noise ability of our model, we further construct a Real-world multimodal NER dataset[1]. The dataset consists of plain text posts and multimodal posts from Twitter, which is more closer to the posts on social platforms. Our method achieves not only state-of-the-art performance but also the superior adaptability and stability in the extensive experiments over both the classical Twitter multimodal NER dataset and our proposed Real-world multimodal NER dataset.

The main contributions of our work are as follows:

(1)We design a Hierarchical Self-adaptation Network (HSN) to identify named entities in multimodal data of social media. It promotes semantic interaction in different modalities and handles the problems that the images are missing or mismatched with the texts, which is more suitable for real-life situations.
(2)To obtain more semantic interaction between different modalities, the HSN employs a Multi-head Hierarchical Attention, which learns the relevant information in different representations subspace iteratively.
(3)We propose a bi-directional manner called Self-adaptive Multimodal Integration to adaptively assign fusion weights for output features, which can solve the situation that images are missing or mismatched with the text.
(4)To evaluate the adaptability and anti-interference ability of the HSN in the real situation, we construct a Real-world multimodal NER dataset.

2. Related work

2.1. Multimodal named entity recognition

There are lots of previous works [11–13] study in short and noisy texts in social media data. However, these methods only utilize the textual part of the post, which is usually insufficient to provide strong evidence for named entity recognition in social media. In recent years some researchers explored the interaction between texts and images from social media posts to improve this task. Zhang et al. [8] proposed a Co-attention model using the simple concatenating operation to fuse textual features and multimodal features. Lu et al. [9] and Arshad et al. [10] also ignored the plain text information, and using an attention mechanism to capture the relevance between image and text. As we know, the multimodal NER is a more text-oriented task, and the image features are auxiliary for better tagging. However, they use single attention to calculate the matching between both features, which may lead to incomplete or wrong attention in multimodal interactivity. Furthermore, they can not solve the plain text posts problem and the mismatch of text and image, and these posts are very common in real-world social media. In our work, we proposed the Hierarchical Self-adaptation Network to solve the insufficient of previous models, namely the lack of visual information in the corpus and incomplete or wrong attention in multimodal interactivity. And when dataset consists of both plain text posts and multimodal posts, i.e., the real-world scenario. Our method can also obtain state-of-the-art performance.

2.2. Multimodal attention mechansim

Multimodal attention allows models to focus on both important parts of images or text of a task. It has been successfully applied to vision and language related tasks. Remi et al. [14] employed a multimodal attention mechanism for VQA to find regions in images that are most related to the text. Xu et al. [15] used for image captioning to generate a word based on the visual area that is most

[1] The Real-world NER dataset in this paper is downloaded from https://github.com/T1aNS1R/Real-world-multimodal-NER-dataset.

related to the last generated word. Our attention implementation approach is similar to those used for VQA, which aims to obtain more semantic interaction between different modalities. The differences are that we use Multi-head Hierarchical Attention to capture important features and fuse them to get multimodal features. We employ a bi-directional integration to fully utilize the allocation of plain textual features and fused features, which can efficiently suppress the noise in fused features, and share more weight on them.

3. Hierarchical self-adaptation network

In this section, we first define the problem formulation, then introduce the details of the Hierarchical Self-adaptation Network (HSN).

3.1. Problem formulation

The MNER task is defined as follows: given a post containing a sentence $S = (s_1, \ldots s_n)$ and an image I as input, the aim is to learn a labeling function to predict the label sequence $\boldsymbol{l} = (l_1, \ldots, l_n)$ for the post.

HSN for the MNER task can be framed as a hierarchical three-layered architecture, as illustrated in Fig. 2. The left panel is the Multimodal Feature Extractor to get textual features $\boldsymbol{H}$ and visual features $\boldsymbol{V}$ from post $\langle S, I \rangle$.

The middle panel is the Cross-modal Interaction Module (CIM) includes *Text-to-Image Unit* and *Image-to-Text Unit*. We use Multimodal Hierarchical Attention in both units to iteratively capture more cross-modal semantic interaction in different representations subspace, suppress incomplete or wrong attention in multimodal interactivity. We obtain multimodal fusion features $\boldsymbol{F}$ by CIM, which benefits from the mutual information between textual features and visual features.

$$\boldsymbol{F} = CIM(\boldsymbol{I}, \boldsymbol{H}) \tag{1}$$

The right panel is the Self-adaptive Multimodal Integration (SMI), which merges textual features $\boldsymbol{H}$ and fusion features $\boldsymbol{F}$ to obtain the multimodal features $\boldsymbol{O}$. It can restrain the noise in fusion features and improve the utilization of useful features.

$$\boldsymbol{O} = SMI(\boldsymbol{H}, \boldsymbol{F}) \tag{2}$$

Then we use a CRF layer as a decoder to assign the sentence in the post a label sequence $\boldsymbol{l}$:

$$p(\boldsymbol{l}|\boldsymbol{O}) = \frac{1}{\boldsymbol{Z}(\boldsymbol{O})} \prod_{t=1}^{N} \psi(l_{t-1}, l_t, \boldsymbol{O}; \theta) \tag{3}$$

$$\boldsymbol{l}^* = argmax_{\boldsymbol{l}' \in \boldsymbol{l}} p(\boldsymbol{l}|\boldsymbol{O}) \tag{4}$$

where $\boldsymbol{Z}(\boldsymbol{O})$ is a normalization term that adds up the products of $\psi(\cdot)$ for all the possible label sequences. $\psi(\cdot)$ is the potential function, with parameters θ, and the maximum a posteriori sequence $\boldsymbol{l}^*$ can be computed using dynamic programming.

The proposed optimization problem will be approved by the following.

3.2. Multimodal feature extractor

3.2.1. Text feature extractor

Our proposed model uses Word-level, Char-level, and POS-level embeddings to get the final representation of each token s_i in the sentence S: 1) pretrained embeddings to initialize Word-level representation $\boldsymbol{x}^w$; 2) Convolutional Neural Networks to randomly initialize Char-level representation $\boldsymbol{x}^c$; 3) embedding matrix to randomly initialize POS-level representation $\boldsymbol{x}^p$. Then we obtain the final representation of each token $\boldsymbol{x}_i = [\boldsymbol{x}^w; \boldsymbol{x}^c; \boldsymbol{x}^p]$ by concatenating three level representations.

Afterwards, we further feed the final representation $\boldsymbol{x}_i$ into highway networks[16] with *tanh* output activation.

$$\tilde{\boldsymbol{h}}_i = tanh(Highway(\boldsymbol{x}_i)) \tag{5}$$

Then we implement BiLSTM [17] to capture the contextual information of each token in the sentence:

$$\boldsymbol{h}_i = BiLSTM\left(\tilde{\boldsymbol{h}}_i\right) \tag{6}$$

where $\boldsymbol{h}_i$ is the ith representation of the textual feature matrix $\boldsymbol{H} \in \mathbb{R}^{n \times d}$, d is the hidden size of BiLSTM.

3.2.2. Image feature eextractor

To obtain the representations of images, we use the pretrained 16-layer VGGNet model [18]. We first choose the representation of

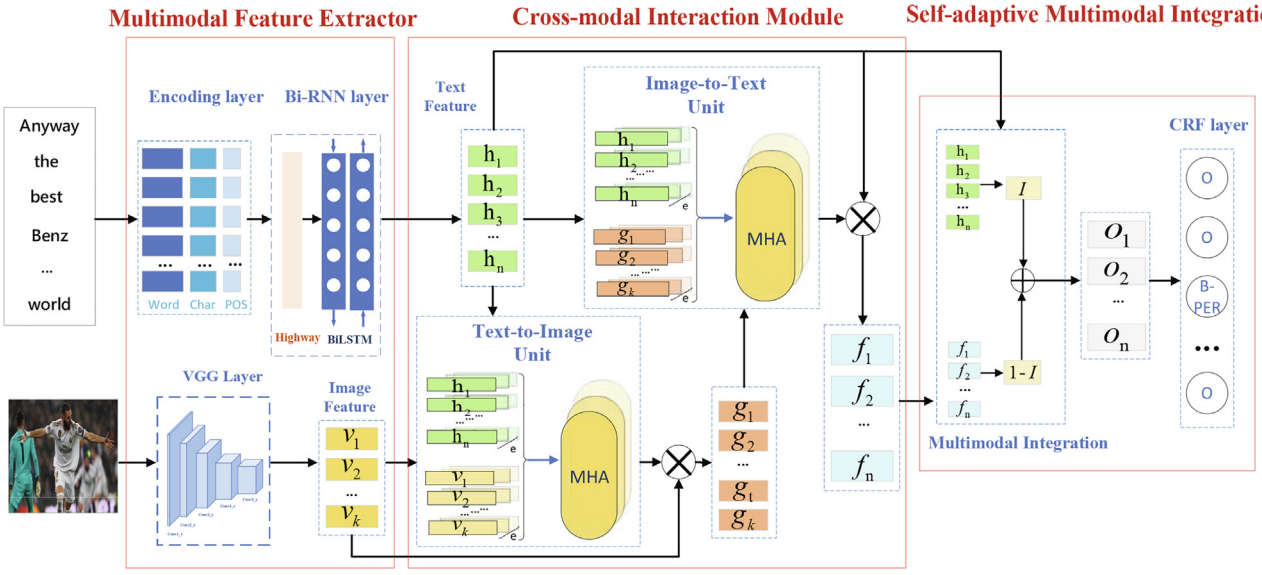

Fig. 2. Overall architecture of the Hierarchical Self-adaptation Network (HSN). The HSN consists of three parts, a multimodal feature extractor, a cross-modal interaction module, and a self-adaptive multimodal integration.

image $\tilde{\boldsymbol{V}}$ from the last pooling layer of VGG16, then use a non-linear layer to transform $\tilde{\boldsymbol{V}}$ with the same dimension with $\boldsymbol{H}$:

$$\boldsymbol{v}_j = relu(W_v \tilde{\boldsymbol{v}}_j + b_v) \tag{7}$$

where W_v, b_v are the weight matrices and the biases respectively, $\boldsymbol{v}_j$ is the *j-th* representation of visual feature matrix $\boldsymbol{V} \in \mathbb{R}^{k \times d}$after non-linear layer, k is the number of regions.

3.3. Cross-modal interaction module

Our proposed model employs Multi-head Hierarchical Attention (MHA) in both units of the module, which is composed of Multimodal hierarchical attention and Multi-head process. It can solve incomplete or wrong cross-modal attention caused by multiple entity objects.

The MHA adopts a hierarchical architecture, which can learn the relevant information in different representations subspace iteratively. The MHA is bi-level attention, different from traditional attention mechanisms, we calculate twice the attention of each feature. For simplicity, we mainly introduce its application in *Text-to-Image Unit* as the following, the process of *Image-to-Text Unit* is similar.

3.3.1. Multimodal hierarchical attention

To obtain more semantic interaction between different modalities and suppress wrong cross-modal attention, the HSN proposes a hierarchical attention mechanism that can learn relevant information iteratively. A graphical representation of Multimodal hierarchical attention is available in Fig. 3. In each cross-modal interaction unit, the multi-modal feature will hierarchically interact twice.

Specifically, we first calculate the attention weight of each element in both representations via dot product:

$$\boldsymbol{M}(i,j) = \boldsymbol{v}_i^T \cdot \boldsymbol{h}_j \tag{8}$$

where $\boldsymbol{M}(i,j)$ represents the value of ith row and jth column calculated by dot product. We can get attention weight matrix for each element in $\boldsymbol{H}$ and $\boldsymbol{V}$.

Similar to work of [19], the probability distributions of *Image-level attention* $\boldsymbol{A}$ is calculated by column-wise softmax function in each column. The tth column represents a *Text-to-Image attention* $\boldsymbol{\alpha}(t)$, when viewing a single word $\boldsymbol{h}_t$ only.

$$\boldsymbol{\alpha}(t) = softmax(\boldsymbol{M}(1,t), \ldots, \boldsymbol{M}(k,t)) \tag{9}$$

$$\boldsymbol{A} = [\boldsymbol{\alpha}(1), \boldsymbol{\alpha}(2), \ldots, \boldsymbol{\alpha}(n)] \tag{10}$$

Like *Image-level attention*, Our model uses a row-wise softmax function to get probability distributions of *Text-level attention* $\boldsymbol{\beta}$. To get more interactive informations from image to text, we average all the $\boldsymbol{\beta}(t)$ to get an averaged *Text-level attention*.

$$\boldsymbol{\beta}(t) = softmx(\boldsymbol{M}(t,1), \ldots, \boldsymbol{M}(t,n)) \tag{11}$$

$$\boldsymbol{\beta} = \frac{1}{n}\sum_{t=1}^{k} \boldsymbol{\beta}(t) \tag{12}$$

At last, we calculate dot product of $\boldsymbol{A}$ and $\boldsymbol{\beta}$ to get the *Final Image-level attention* $\boldsymbol{s}$. In this way, the model first calculate the importance of each word in the sentence, then getting the weight of each word to the importance pairs of each visual feature $\boldsymbol{v}$.

$$\boldsymbol{s} = \boldsymbol{A}^T \boldsymbol{\beta} \tag{13}$$

Different from [19], we make significant changes to its overall, which structure is capable of satisfying cross-modal interaction with heterogeneous features. Through this hierarchical approach, the MHA can fully learn the information of each modal.

3.3.2. Multi-head process

Inspired by [20], Our proposed model employs the Multi-head process to enhance Multimodal hierarchical attention. It enables our attention mechanism to capture multiple important information in different representation subspaces, rather than only focus on local information. So our model can solve incomplete or wrong cross-modal attention caused by multiple entity objects. The detailed structure of the Multi-head process as shown in Fig. 4.

At first, the features pass of a liner layer, then enter output into the Text-to-Image attention, we do this operation e times. This operation can learn relevant information in different representation subspaces.

$$MultiHead(\boldsymbol{V}, \boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{|e|} softmax(head_i) \tag{14}$$

$$where\ head_i = Atten\left(\boldsymbol{V}W_i^v, \boldsymbol{H}W_i^{h_1}\right) \tag{15}$$

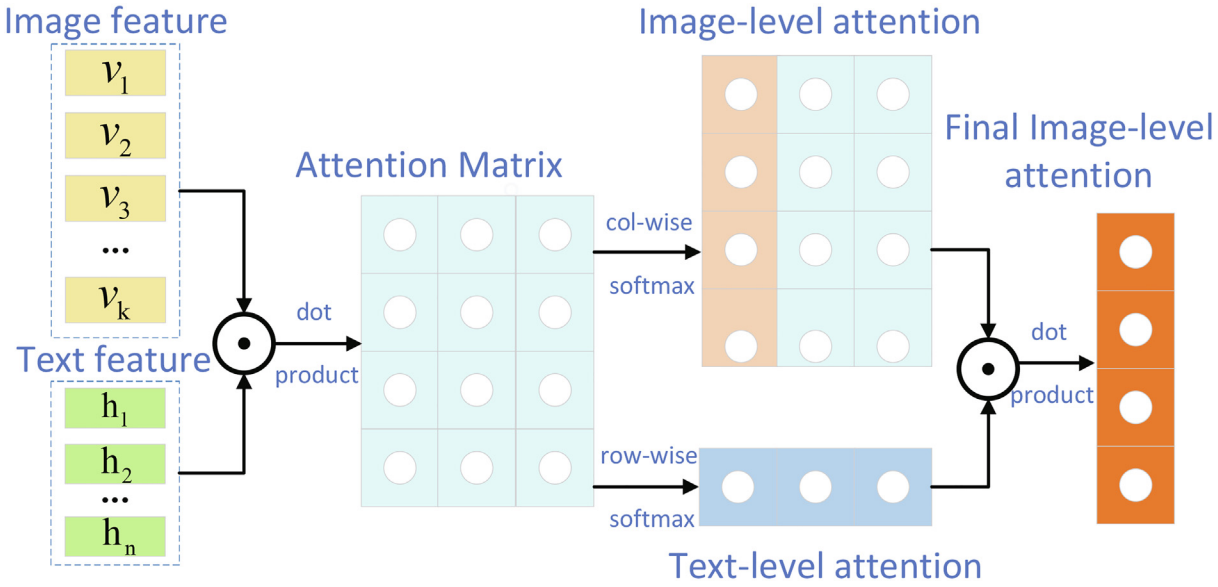

Fig. 3. The detailed structure of multimodal hierarchical attention, taking the text-to-Image unit for example.

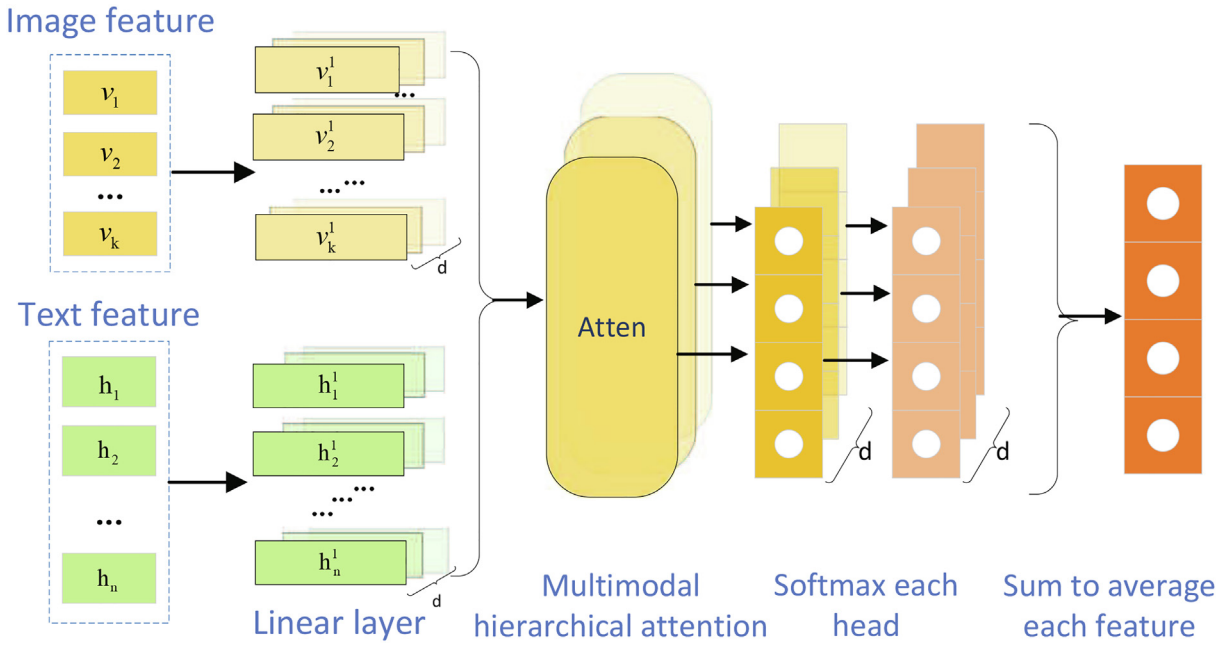

Fig. 4. The detailed stucture of multi-head process, taking the text-to-Image unit for example.

Different from [20], the model first apply softmax on each head to get the weight of the representation subspace, then sum to average each head to get final Text-to-Image attention weight $\boldsymbol{s}_{img}$, which has learned more important information than single attention. If we simply use multi-head attention in multimodal interaction, most of the weights will be replayed on the core object, suppressing other potential objects. We make substantial changes to the internal structure to solve incomplete or wrong cross-modal attention problems caused by multiple objects.

Then we get the text-guided image representation $\boldsymbol{G}$ by element-wise product function $\boldsymbol{s}_{img}$ and $\boldsymbol{V}$ as follow:

$$\boldsymbol{G} = \boldsymbol{s}_{img} \odot \boldsymbol{V} \tag{16}$$

We use the same attention mechanism in *Image-to-text unit*. $\boldsymbol{H}$ and $\boldsymbol{G}$ are input for MHA and get *Final Text-level attention* $\boldsymbol{s}_{sen}$, the multimodal fusion representation $\boldsymbol{F}$ is calculated by element-wise product function $\boldsymbol{s}_{sen}$ and $\boldsymbol{H}$.

3.4. Self-adaptive multimodal integration

Considering when the image is mismatched with the text, the fusion representation $\boldsymbol{F}$ will carry irrelevant noise information and bring interference to our label prediction. To solve this problem, we merge textual features $\boldsymbol{H}$ and fusion features $\boldsymbol{F}$ with the integration to restrain the noises in fusion features, which could regulate the proportion of output by itself. The integration γ consists n scalars in range of $[0, 1]$. In an extreme case, if the image is not at all relevant to the text, the value of each scalar is 1. Conversely, the value of each scalar is 0. The integration function is defined as follow:

$$\gamma = \sigma(W_\gamma \boldsymbol{H} + U_\gamma \boldsymbol{F} + b_\gamma) \tag{17}$$

$$\boldsymbol{O} = \gamma \odot \boldsymbol{H} + (1 - \gamma) \odot \boldsymbol{F} \tag{18}$$

where W_γ, U_γ are self-adaptive weight matrices, $b_\gamma$is the biases, σ is the element-wise sigmoid function, $\boldsymbol{O}$ is the final feature.

Different from Zhang et al. [8] and Lu et al. [9], SMI uses a bi-directional scheme to utilize correlation in fusion features and textual features fully. Compared with the previous gate scheme, i.e., single direction aided manner. Our scheme can share more weight on fusion features, which can efficiently capture the important cross-modal semantic interaction.

Algorithm 1: HSN

Input: sentence and image tuple $\langle S, I \rangle$
Output: Labels L of the sentence
Step1: Multimodal Feature Extractor
  $\boldsymbol{H} \leftarrow TextFeatureExtractor(S)$
  $\boldsymbol{V} \leftarrow ImageFeatureExtractor(I)$
Step2: Cross-modal Interaction
  Text-to-Image weight $\boldsymbol{s}_{img} \leftarrow MHA(\boldsymbol{H}, \boldsymbol{V})$
  Text-guided Image feature $\boldsymbol{G} \leftarrow \boldsymbol{s}_{img} \odot \boldsymbol{V}$
  Image-to-Text weight $\boldsymbol{s}_{sen} \leftarrow MHA(\boldsymbol{G}, \boldsymbol{H})$
  multimodal fusion feature $\boldsymbol{F} \leftarrow \boldsymbol{s}_{sen} \odot \boldsymbol{H}$
Step3:Self-adaptive Multimodal Integration
  The final feature $\boldsymbol{O} \leftarrow Integration(\boldsymbol{H}, \boldsymbol{F})$
  The labels $L \leftarrow CRF(\boldsymbol{O})$

At last, we add a CRF layer inspired by [21]. CRFs [22] has been successfully used in sequence labeling tasks which have the important constraints of the labels (e.g., I-PER must follow B-PER).

The workflow of our method can be seen in Algorithm 1.

4. Experiment

To validate the effectiveness of our proposed method, we perform extensive experiments on the Real-world multimodal NER dataset and the Twitter multimodal NER dataset. In this section, we first introduce the datasets and baselines, then we perform the comparative experiment and the ablation experiments to validate the efficiency of the proposed model. Finally, we report the qualitative analysis to show the interpretability of our attention modules. Precision, Recall, and F1 are used as evaluation metrics in our experiments. A named entity is considered correctly recognized only if its both boundaries and type match ground truth.

4.1. Datasets

We concisely introduce datasets used in our experiments, and the data statistics of them are shown in Table 1. There are four types of named entities (*Person, Organization, Location, Miscellaneous*) in the datasets.

4.1.1. Twitter multimodal NER dataset

We conduct our model on Twitter multimodal NER dataset (TMN dataset) [8] collected from tweets, and each data instance consists of a pair of a sentence and an image.

4.1.2. Real-world multimodal NER dataset

To evaluate the adaptability of our proposed approach in the real-life scenario. We construct a Real-world multimodal NER dataset. The dataset is based on the TMN dataset and added lots of plain text posts to make it closer to the real social platform. We extract sentences from Twitter using Twitter's API, and messages without named entities will be dropped out. We add 3000/1000/2000/ tweets to expand the TMN dataset. We use the standard BIO schema in this work, and the topics covered are diverse in nature, such as *sports, music, social event*. The new messages are labeled by three independent expert annotators.

We further analyze the properties of the Real-world multimodal NER dataset. There are 1679 images include multiple subjects, and we find when more than three persons in images, the probability of *organization* appearing in the text is higher than *person*. The size of noisy images (blur, table, or poor alignment with text.) is 1029.

4.2. Baselines

In this subsection, we describe the comparisons of our main experiments. We report the performance of the following state-of-the-art NER model, with variations of our proposed approach to examine contributions of each component.

- **CNN + BiLSTM + CRF** [21]: An End-to-end system benefits from both word- and character-level representations automatically by using a combination of bidirectional LSTM, CNN, and CRF. And it has achieved the best result on the CoNLL 2003 test set.
- **Adaptive Co-Attention Network** [8]: A multimodal method leverage the co-attention process to capture the semantic interaction between different modalities. And their gate scheme only works on fused features, which will then be appended to textual features.
- **Disan MultiModal NER** [10]: Previous state-of-the-art MNER model inspired by Directional self attention network (Disan) [23], which extends multi-dimensional self attention approaches to jointly learn intra and cross-modal dependencie.
- **(proposed) HSN** is the proposed approach, as described in Fig. 2. The HSN consists of three parts, a multimodal feature extractor, a cross-modal interaction Module, and a self-adaptive multimodal integration.
- **(proposed) w/o Multimodal Integration (MI)** is a part of our method without Multimodal Integration. It directly concatenates both heterogeneous features instead.
- **(proposed) w/o Hierarchical Attention (HA)** is a variant of our approach that uses Luong Attention [24] instead of Multimodal Hierarchical Attention.
- **(proposed) w/o Multi-head process (MP)** is a variant of our model without Multi-head process.
- **(proposed) w/o visual feature (VF)** To verify whether the visual features can assist in better labeling, we use a blank picture tiled with default constant instead of the accompanying image.

Table 1
The statistics of both multimodal NER datasets.

		Train	Dev	Test
Twitter	Sentences	4000	1000	3527
	Images	4000	1000	3527
	Entities	6176	1546	5072
Real-world	Sentences	7000	2000	5527
	Images	4000	1000	3527
	Entities	11876	3418	9791

4.3. Parameter settings

During the training, the model achieves the best performance by using Adam optimizer with learning rate initialization is 0.003, batch size and dropout rate are 16 and 0.5 on both datasets. We apply early stopping with the patience of 20 to avoid the model from over-fitting. Furthermore, we summarize the hyper-parameters in Table 2. All of the models are implemented with Keras and conducted on a single Nvidia P100 graphic card.

For POS-level representation in Text Feature Extractor, we use the NLTK [25] toolkit to extract the part-of-speech of each word in sentences. We divide part-of-speech into the following ten categories: *Adjective, Verb, Noun, Preposition or subordinating conjunction, Coordinating conjunction, Symbol, Wh-pronoun, Adverb, Determiner, and Other*. We construct a part-of-speech vocabulary based on the above categories and randomly initialize and learn the POS-level representation during training.

For a fair comparison, we use two different word embeddings to compare with previous works to analyze the effect of our methods. One is Twitter Embeddings [8] trained on 30 million tweets, the other is 300D Crawl Embeddings [26] trained on 600B tokens. And the out-of-vocabulary (OOV) words are initialized by Gaussian random sampling. When the post does not contain an image, we use a blank picture tiled with default constant as the input.

4.4. Quantitative analysis

4.4.1. Model comparison on twitter multimodal NER dataset

Table 3 shows the overall results of various methods on the TMN dataset. The first part is the results of models using the Twitter Embeddings, and the second part is the main results of models utilizing the Crawl Embeddings. Our final model **HSN** achieves the best F1 score, establishing a solid state-of-the-art result.

Compared to previous state-of-the-art models on the TMN dataset, our model achieves the best results using both embeddings. The F1 score on our approach is improved by 1.27% in the TMN dataset, which proves the advantage of our model in multimodal NER tasks. The improvement comes from the following: 1) HSN can obtain more semantic interaction between different modalities, and reduce the wrong attention between text and image; 2) our methods could assign more weights to textual features by using self-adaptive multimodal integration when the image mismatch with the text.

Table 2
Hyper-parameters of the modal.

Hyper-parameters	Value
Char-level embedding size	50
POS-level embedding size	50
Image feature size	400
BiLSTM hidden state size	400
parallel attention head	8

Table 3
The result of our approch with comepeting methods on the TMN dataset. The first part is the results of models using the Twitter Embeddings, and the second part is the main results of models utilizing the Crawl Embeddings.

Methods	PER F1	LOC F1	ORG F1	MISC F1	Overall F1
CNN + BiLSTM + CRF	71.25	67.73	38.33	21.98	67.15
Adaptive Co-Attention	81.98	**78.95**	53.07	34.02	70.69
Network	82.38	78.22	55.88	33.00	71.55
Disan MultiModal NER	84.24	78.62	58.91	35.26	72.52
HSN					
Disan MultiModal NER	83.98	78.65	59.27	39.54	72.91
HSN	**85.87**	78.63	**61.93**	**42.74**	**74.18**

Table 4
The result of our approch with comepeting methods on the Real-world dataset. All models utilizing the Crawl Embeddings.

Methods	PER F1	LOC F1	ORG F1	MISC F1	Overall F1
CNN + BiLSTM + CRF	85.27	78.42	**62.52**	41.96	73.96
Adaptive Co-Attention	83.92	78,31	59.31	36.54	72.41
Network	84.25	77.52	60.17	38.64	72.80
Disan MultiModal NER	**86.37**	**79.60**	62.37	**42.39**	**74.92**
HSN					

Table 5
Ablation results of HSN on F1 scores in TMN dataset.

Twitter	Prec.	Recall	F1
w/o MI	73.49	72.56	73.02
w/o HA	73.25	71.69	72.46
w/o MP	74.53	72.92	73.71
w/o VF	72.25	71.97	72.11
HSN	**74.92**	**73.45**	**74.18**

The F1 scores for entities "PER", "ORG", and "MISC" on our approach are all improved, and the score of "LOC" category is almost consistent with the previous state-of-the-art model. And it is interesting to note that entities "PER", "MISC" have a clear performance boost of about 2–3%. We believe this case is due to the greater impact of image features on these entity categories, and our methods can fully utilize more useful information from image features to assist textual features for tagging tasks.

4.4.2. Model comparison on real-world dataset

The results on the Real-world dataset are shown in Table 4, we collect several experiment findings from the results.

The experiments reveal the performance of previous MNER models is worse than the traditional NER model, demonstrating previous models are disturbed by plain text posts. When the dataset contains plain text posts, i.e., closer to the real social platform, the previous models do not perform well. Compared to all previous models, our final model can get the best F1 scores on the Real-world datasets, proving the self-adaptive multimodal integration can realize the autonomic regulation between modules.

Experiments on the Real-world datasets show our model has superior adaptability and anti-interference ability on real data, when the post without the image, the self-adaptive multimodal integration can share more weight on textual features. The results confirm our model can adapt to real-life situations and achieve excellent performance.

4.4.3. Analysis on variant models

We conduct some ablation experiments in our final model, and the results are shown in Table 5. Each component of our model plays a key role in improving performance.

First, we remove the Multimodal Integration and directly concatenate textual features and image features. This results in a loss of 1.16% F1 score on the test set, showing the integration can significantly improve our model by controlling the allocation of both features. Our integration uses a bi-directional scheme to fully utilize correlation in fusion features and textual features, which can efficiently suppress the noise in fused features and share more weight on them.

Then, we adopt Luong Attention [24] to substitute Multimodal Hierarchical Attention. The F1 score descends by 1.72% compared to our model. Results show that our hierarchical attention mechanism can learn more relevant information iteratively between different modalities.

We also employ a single Hierarchical Attention to replace multi-head Hierarchical Attention. The performance drops 0.47%, indicating that the multi-head process can solve incomplete or wrong cross-modal attention problems caused by multiple objects in both features.

Lastly, we use a blank picture tiled with a default constant instead of an accompanying image. The F1 score falls by 2.07% compared to the performance of our model. Results show the visual features could help our model to find the entities and validate entity types. Like Fig. 1, when there is a football player in the image, the sentence is more likely to contain a person's name, but when there is a car in the image, the sentence is more likely to contain a brand of the car.

4.5. Qualitative analysis

4.5.1. Visualization of fully aligned posts

We visualize some fully aligned posts as examples in Fig. 5. The top part is attention visualization for text-to-image, the middle part is the heat maps of image-to-text, and the bottom part is the visualization of the multimodal integration γ where the darker color represents the textual feature is the main contribution.

As the 5(a) illustrates, a single entity object matching between the text and the image. A player on the gym in Fig. 5(a), and our model correctly focuses on the player in the image and *Derek Jeter* in the sentence. The integration γ is more inclined to fusion features when the word in sentence match with the regions of the image.

Fig. 5(b) shows that our model also can perform well when there are multiple entity objects are matching between the text and the image. Our model matches the man and the building properly in both modalities, and the integration γ assigns more weights to fusion features on *Helmut Kohl* and *Brandenburg gate* in the sentence.

4.5.2. Visualization of confusing posts

Fig. 6 shows some confusing examples that trouble the Cross-modal Interaction Module in our model. The layout of Fig. 6 is the same as Fig. 5.

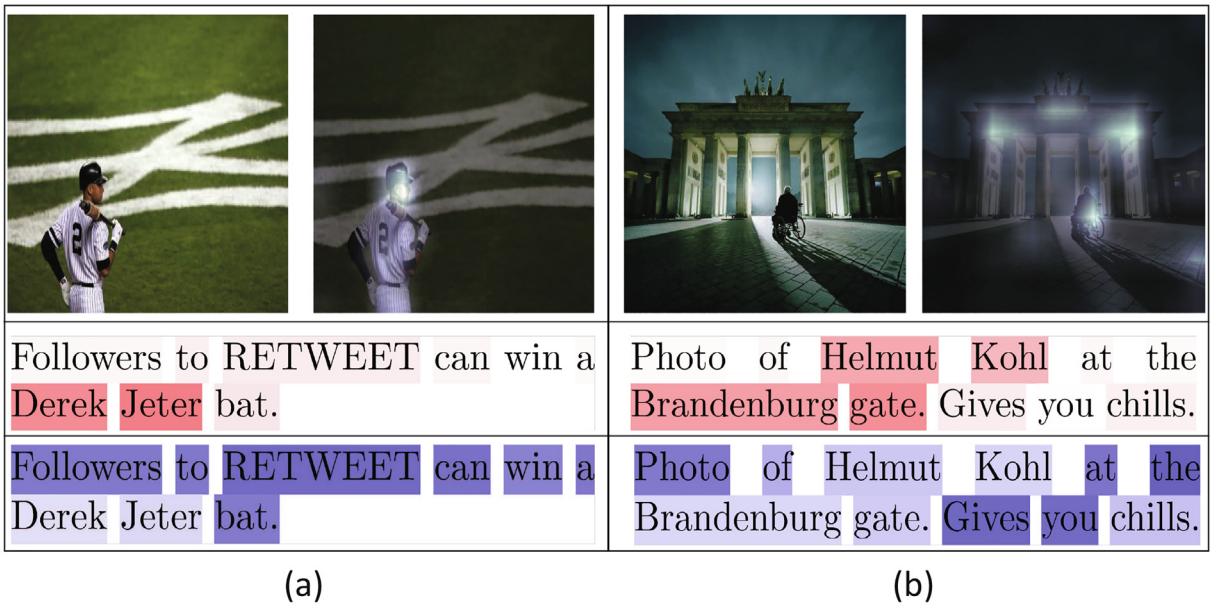

Fig. 5. Some fully aligned examples of visualization in our model. The top part is attention visualization for text-to-image, the middle part is the heat maps of image-to-text, and the bottom part is the visualization of multimodal integration γ. The darker color represents a higher weight.

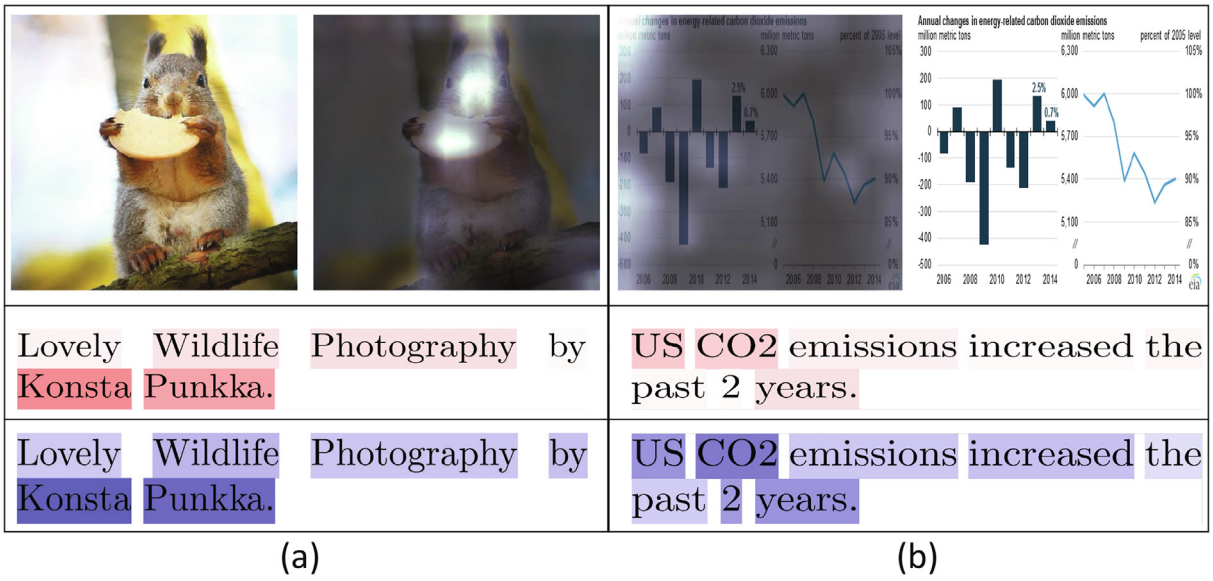

Fig. 6. Some confusing examples that trouble the Cross-modal Interaction Module in our model.

On the one hand, MNER greatly benefits from visual features when the text and the image are well aligned, but this is not always the case in social media. In Fig. 6(a), the sentence shows who is the photographer of the image, but our attention mechanism still focuses on a squirrel in the image and *Konsta Punkka* in the sentence. It gets wrong cross-modal attention in different modalities, but the multimodal integration can solve this problem by assigning more weights on textual features.

On the other hand, images contain more than objects or scenes, and sometimes they can also include text or tables. When there is a table or text in the picture, even if the sentences are well aligned with the relevant image, our model could not get any information from the image. Like 6(b), there is a table about the growth of U.S. CO2 emissions in recent years in the image, it is related to the sentence, but the attention mechanism only extract the noise from visual information. Our proposed model can employ multimodal integration to share more weights on textual features.

The results show our model can successfully focus on appropriate features when the text is matching with the image, and when the image mismatch with the text or the post without the image, our model can assign more weights on textual features by the multimodal integration γ.

5. Conclusions and future work

In this paper, we reveal the essential mismatching problem in the utilization of both textual information and image information in the MNER task. We propose a Hierarchical Self-adaptation Network to identify named entities in multimodal posts, which could use more textual information and suppress incomplete or wrong attention in multimodal interactivity. We employ Multi-head Hierarchical Attention, which learns more semantic interactivity in different representations subspace. Furthermore, we adopt a

Self-adaptive multimodal integration to restrain the noises in fusion features and share more weight on them.

As we know, there are both plain text posts and multimodal posts on social platforms. We believe an excellent model should be able to process these two kinds of content accurately at the same time, so as to apply to real-life social media. To evaluate the adaptability and anti-interference ability of our model, we construct a Real-world multimodal dataset consisting of plain text posts and multimodal posts from Twitter.

We conduct extensive experiments on both the classical Twitter multimodal NER dataset and the constructed datasets. As expected, the experimental results in both datasets validate the efficiency of the proposed model with state-of-the-art performances, especially in the case of the images are missing or mismatched with the texts. The ablation experiments reflect the importance and effectiveness of each component in our model, and the qualitative analysis shows the interpretability of our attention modules.

We hope our model can significantly inspire scholars in multimodal social media and encourage more research on multimodal social media to apply to real-life situations. In the future, we would like to make up the omissions in current work and expand our model to more multimodal tasks such as Entity Liking and fine-grained name tagging.

CRediT authorship contribution statement

Yu Tian: Conceptualization, Methodology, Writing - original draft, Software. **Xian Sun:** Writing - review & editing, Validation, Supervision. **Hongfeng Yu:** Data curation, Formal analysis. **Ya Li:** Visualization, Investigation. **Kun Fu:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] I. Calixto, M. Rios, W. Aziz, Latent variable model for multi-modal translation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6392–6405.

[2] I. Calixto, M. Rios, W. Aziz, Latent variable model for multi-modal translation, in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28– August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 6392–6405. doi:10.18653/v1/p19-1642. https://doi.org/10.18653/v1/p19-1642..

[3] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J.M. Moura, D. Parikh, D. Batra, Visual dialog, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 326–335.

[4] D.A. Hudson, C.D. Manning, Gqa A new dataset for real-world visual reasoning and compositional question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6700–6709.

[5] P.P. Liang, A. Zadeh, L. Morency, Multimodal local-global ranking fusion for emotion recognition, in: S.K. D'Mello, P.G. Georgiou, S. Scherer, E.M. Provost, M. Soleymani, M. Worsley (Eds.), Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16–20, 2018, ACM, 2018, pp. 472–476. doi:10.1145/3242969.3243019. doi:10.1145/3242969.3243019..

[6] P.P. Liang, Z. Liu, A. Zadeh, L. Morency, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Multimodal language analysis with recurrent multistage fusion Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 150–161, https://doi.org/10.18653/v1/d18-1014.

[7] T. Baldwin, M.-C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, W. Xu, Shared tasks of the 2015 workshop on noisy user-generated text: twitter lexical normalization and named entity recognition, in: Proceedings of the Workshop on Noisy User-generated Text, 2015, pp. 126–135..

[8] Q. Zhang, J. Fu, X. Liu, X. Huang, Adaptive co-attention network for named entity recognition in tweets, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018..

[9] D. Lu, L. Neves, V. Carvalho, N. Zhang, H. Ji, 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), 2018, pp. 1990–1999..

[10] O. Arshad, I. Gallo, S. Nawaz, A. Calefati, Aiding intra-text representations with visual context for multimodal named entity recognition, arXiv preprint arXiv:1904.01356..

[11] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: K. Knight, A. Nenkova, O. Rambow (Eds.), NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, The Association for Computational Linguistics, 2016, pp. 260–270. doi:10.18653/v1/n16-1030. https://doi.org/10.18653/v1/n16-1030.

[12] G. Aguilar, S. Maharjan, A.P. López-Monroy, T. Solorio, A multi-task approach for named entity recognition in social media data, in: Proceedings of the 3rd Workshop on Noisy User-generated Text, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 148–153. doi:10.18653/v1/W17-4419. https://www.aclweb.org/anthology/W17-4419..

[13] G. Aguilar, A.P. López-Monroy, F.A.G. Osorio, T. Solorio, Modeling noisiness to recognize named entities using multitask neural networks on social media, in: M.A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 1401–1412. doi:10.18653/v1/n18-1127. https://doi.org/10.18653/v1/n18-1127..

[14] R. Cadène, H. Ben-younes, M. Cord, N. Thome, MUREL: multimodal relational reasoning for visual question answering, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/ IEEE, 2019, pp. 1989–1998. doi:10.1109/CVPR.2019.00209. http://openaccess.thecvf.com/content_CVPR_2019/html/Cadene_MUREL_Multimodal_Relational_Reasoning_for_Visual_Question_Answering_CVPR_2019_paper.html..

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[16] R.K. Srivastava, K. Greff, J. Schmidhuber, Highway networks, CoRR abs/1505.00387. arXiv:1505.00387. http://arxiv.org/abs/1505.00387.

[17] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, N.A. Smith, Transition-based dependency parsing with stack long short-term memory, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 334–343..

[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015. http://arxiv.org/abs/1409.1556..

[19] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, Attention-over-attention neural networks for reading comprehension, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, 2017, pp. 593–602. doi:10.18653/v1/P17-1055. https://doi.org/10.18653/v1/P17-1055.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008..

[21] X. Ma, E.H. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 1: Long Papers, The Association for Computer Linguistics, 2016. https://www.aclweb.org/anthology/P16-1101/..

[22] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data..

[23] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, C. Zhang, Disan: directional self-attention network for rnn/cnn-free language understanding, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018..

[24] T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, in: L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, Y. Marton (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015, The Association for Computational Linguistics, 2015, pp. 1412–1421.

[25] E. Loper, S. Bird, Nltk: the natural language toolkit, arXiv preprint cs/0205028..

[26] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

Yu Tian received the B.Sc. degree from Dalian Maritime University, Dalian, China, in 2018. He is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, pattern recognition, natural language processing and remote sensing image processing, especially on instance segmentation.

Xian Sun received the B.Sc. degree from Beihang University, Beijing, China, in 2004, and the M.Sc. and Ph. D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2006 and 2009, respectively.
He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote-sensing image understanding.

Hongfeng Yu received the B.Sc. degree and M.Sc degree from Peking University, Beijing, China, in 2013 and 2016 respectively.
He is currently a Research Assistant at the Institute of Electronics, Chinese Academy of Sciences. His research interests include deep learning and natural language processing.

Ya Li received the B.Sc. degree and M.Sc. degree from Hebei University of Technology, Tianjin, China, in 2014 and 2017 respectively.
She is currently a engineer at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. Her research interests include deep learning and multimodal information mining.

Kun Fu received the B.Sc., M.Sc., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively.
He is currently a Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining, and visualization.