Object-Aware Multimodal Named Entity Recognition in Social Media Posts With Adversarial Learning

Changmeng Zheng, Zhiwei Wu, Tao Wang, Yi Cai, *Member, IEEE*, and Qing Li, *Member, IEEE*

***Abstract*—Named Entity Recognition (NER) in social media posts is challenging since texts are usually short and contexts are lacking. Most recent works show that visual information can boost the NER performance since images can provide complementary contextual information for texts. However, the image-level features ignore the mapping relations between fine-grained visual objects and textual entities, which results in error detection in entities with different types. To better exploit visual and textual information in NER, we propose an adversarial gated bilinear attention neural network (AGBAN). The model jointly extracts entity-related features from both visual objects and texts, and leverages an adversarial training to map two different representations into a shared representation. As a result, domain information contained in an image can be transferred and applied for extracting named entities in the text associated with the image. Experimental results on Tweets dataset demonstrate that our model outperforms the state-of-the-art methods. Moreover, we systematically evaluate the effectiveness of the proposed gated bilinear attention network in capturing the interactions of mutimodal features visual objects and textual words. Our results indicate that the adversarial training can effectively exploit commonalities across heterogeneous data sources, which leads to improved performance in NER when compared to models purely exploiting text data or combining the image-level visual features.**

***Index Terms*—Named entity recognition, social media posts, adversarial training, bilinear attention network.**

I. Introduction

SOCIAL media like Twitter and Instagram provide massive user-generated information. These information plays a vital role in understanding human behaviors. Social media data are primary in unstructured forms, such as in free-form text or images. Named Entity Recognition (NER), a task which tries to locate and classify named entities into predefined categories, is a critical step in mining social media data and a pre-processing step for many downstream applications, such as breaking news aggregation [1], cyber-attack detection [2] and disease outbreak mining [3].

Traditional neural-based NER models achieve great success in newswire domain but suffer from sharp performance decline in social media domain. The main reason is that texts in social media are usually short and informal, lack of contexts and full of ambiguous expressions [4]. To tackle this problem, many methods are proposed to identify entities with external knowledge base. Ritter *et al.* [5] proposed a distantly supervised method which leverages a large amount of unlabeled data in addition to large dictionaries. Li *et al.* [4] introduced an iterative method to split tweets into meaningful segments and evaluated the method on the NER task. However, these text-based methods purely rely on text data and cannot effectively identify named entities and their types when lacking of textual context. For example, "Charlie" is incorrectly extracted as a person's name rather than a dog's name in the sentence "Charlie is ready for this winter" commenting on an image of a dog. Thus, named entities can be identified not only based on the text data but also visual posts associated with the texts.

Currently, with the development of the deep learning and representation learning, neural network based multimodal NER methods [6]–[9] have been proposed to utilize both image and text information for predicting named entities in social media. Despite showing great improvement against text-based methods, these methods have two limitations:

The first remarkable limitation is that they ignore the mapping relations between visual objects and named entities. In a sentence which contains multiple entities with multiple types, there are more than one mapping between named entities and objects within images. Previous multimodal NER methods [6], [7], [9] representing the image with only one vector which is trained on one semantic label will mislead their models to extract different types of entities into the same type. For example, in Fig. 1, the sentence contains two entities with two different types: a PER entity and a MISC entity. The detected visual object with label "person" is related to the PER entity "Ang Lee". The object "trophy" is related to the MISC entity "Oscars". Previous work will incorrectly extract the two different types of entities in to a same type. Compared to these methods combining image-level features into multimodal features, object-level features (the

Manuscript received March 18, 2020; revised June 16, 2020 and July 24, 2020; accepted July 24, 2020. Date of publication August 3, 2020; date of current version July 30, 2021. This work was supported in part by the Fundamental Research Funds for the Central Universities, SCUT under Grants 2017ZD048 and D2182480, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2017B050506004, in part by the Science and Technology Programs of Guangzhou under Grants 201704030076, 201802010027, and 201902010046 and the collaborative research grants from a CUHK Research Committee Funding (Direct Grants) (Project Code: EE16963) and the Hong Kong Research Grants Council (Project no. C1031-18 G). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. P. K. Atrey. *(Changmeng Zheng and Zhiwei Wu contributed equally to this work.) (Corresponding author: Yi Cai.)*

Changmeng Zheng, Zhiwei Wu, and Yi Cai are with the School of Software Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: sethecharm@mail.scut.edu.cn; zhiwei.w@qq.com; ycai@scut.edu.cn).

Tao Wang is with the Department of Biostatistics and Health Informatics, King's College London, London WC2R 2LS, U.K. (e-mail: wtgmme@gmail.com).

Qing Li is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: itqli@cityu.edu.hk).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.3013398

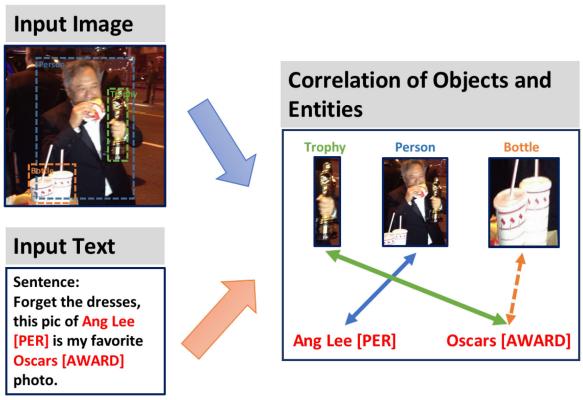


Fig. 1. An example of the Twitter dataset. The visual object with label "person" will lead to the detection of "Ang Lee" as PER category, and objects with "trophy" will lead to the extraction of "Oscars" as the name of an award (MISC). The object "bottle" is irrelevant to entities in this post.

representations of visual objects "person" and "trophy") can reflect the mapping relations of visual objects and textual words. These mapping relations help model to distinguish entities with different types and extract entities precisely. Thus, it is necessary to utilize the object-level features for identifying entities rather than image-level features.

Another limitation is that previous works ignore the distribution disparity of image and text features. They simply concatenate the representations of words and a related image to predict entity labels [6], [7]. Due to the different distributions of image and text, the alignments between named entities and image regions cannot be captured in their methods, resulting in poor NER performance. Therefore, an effective method should be derived to bridge the distribution gaps for robust multimodal representations in social media NER task.

To address the above problems, we propose an adversarial gated bilinear attention network (AGBAN) to better exploit visual and textual information in social media NER. To capture the mapping relations between visual objects and textual entities (the first limitation), we introduce a gated bilinear attention network (GBAN). Bilinear attention network (BAN), which is first proposed by kim *et al.* [10], can exploit the interactions between two groups of input channels, which can be utilized to capture the relations of entities and objects. In Fig. 1, our bilinear attention network can take the mapping relations of different visual objects ("person, trophy") and textual entities ("Ang Lee, Oscars") into account. The different visual objects help to extract entities with different types ("person" for "PER" and "trophy" for "AWARD"). We have also observed that some visual objects are irrelevant to any entities, for example, the visual object "Bottle" is irrelevant to any entities in the sentence in Fig. 1. In that case, we further extend the naive bilinear attention with a object-level gated mechanism which can filter out irrelevant visual objects. Compared to the previous work [6], which filters out the irrelevant visual information in rough image regions, our model precisely extracts the fine-grained objects and improves the NER performance. To address the problem of distribution disparity of image and text features (the second limitation), we

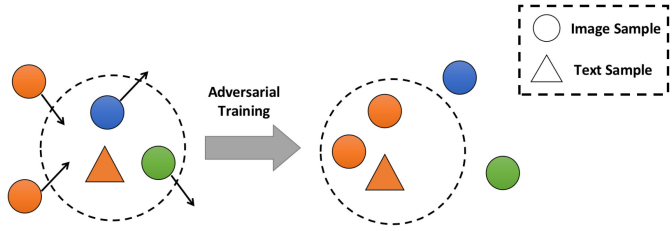


Fig. 2. General idea of our proposed network to achieve an improved, modality-invariant subspace embedding with adversarial training. Shapes of the same color are semantically similar.

propose to extend the GBAN framework with adversarial learning. Motivated by [11], [12], we consider to achieve an improved, modality-invariant subspace embedding with adversarial learning, as shown in Fig. 2. Different from other methods which aim to align single image and text, adversarial training can generate a shared and robust representation in modality space. The modality-invariant subspace is built for a better fusion of the two modalities without the semantic disparity. The fusion representations of visual objects and textual words are sent to a Conditional Random Fields (CRF) layer to predict entity labels.

Our main contributions can be summarized as following:

- We propose an adversarial bilinear attention network to capture the correlations of visual objects and textual entities. The adversarial learning can build a common subspace for a better fusion of the two modalities. The bilinear attention network helps to align entities with related visual objects. To the best of our knowledge, this paper is the first to introduce adversarial learning and bilinear attention mechanism in multimodal named entity recognition task to solve the problem of modality gaps and image-text alignments.
- We extend the bilinear attention network with an object-level gated mechanism to filter out the irrelevant visual objects. Our model has a better performance than previous image-level multimodal NER methods as it considers the fine-grained visual features and precisely filters out irrelevant objects instead of rough image regions.
- We evaluate our method on multimodal named entity recognition dataset based on Twitter data. Our experimental results show that (i) the visual object information from social media posts can improve the performance of NER significantly, and (ii) the adversarial bilinear attention network achieves a better representation for entity-object alignments.

II. Related Work

A. NER in Social Media

Named Entity Recognition (NER) has drawn attention of Natural Language Processing (NLP) researchers because it is a fundamental task in information extraction and can significantly influence the performance of many downstream NLP tasks such as relation extraction [13] and entity linking [14]. Neural models have been proposed and achieved the state-of-the-art performance in various datasets and domains [15]–[18]. Recently,

NER in social media domain has raised attention since texts in social media are explosively growing and provide abundant user-generated information for various applications such as the identification of natural disasters [19], [20], cyber attack detection [2], [21] and breaking news aggregation [22]. However, these NER methods suffer from highly reduced performance deterioration on social media posts because of the noisy and short nature of texts on social media.

There have been several models to improve the traditional NER framework for a better performance on social media. Gimple *et al.* incorporate tweet-specific features including atmentions, hashtags, URLs, and emotions obtained by using a new labeling scheme [23]. Ritter *et al.* propose a T-NER system which uses LabeledLDA to exploit Freebase dictionaries as a source of distant supervision [5]. However, their method only identifies whether a span is an entity or not. Recently, Baldwin *et al.* and Aguilar *et al.* report that performance gains from leveraging external sources of information such as lexical information (e.g., POS tags, etc.) and/or from several preprocessing steps (e.g., token substitution, etc.) [24], [25].

Most of the previous methods only take the textual content into account. However, people like to share their daily lives in social media using not only texts but also image posts related to the texts. Such visual content can help us recognize named entities. Moon *et al.* [7] and zhang *et al.* [6] propose to leverage the visual information to help extract entities. The visual content and textual representations are related by attention mechanism. However, they represent an image with one single vector trained with only one semantic label, which cannot assist recognizing multiple entities with different types. Our model introduces object level representations to make the attention more focus on effective regions in images and entity-relative objects can help to extract the entities more precisely.

B. Multimodal Representation

A large number of researches have shown that combining textual and visual representation as multimodal features can improve the performance of semantic extraction tasks [26]–[29]. Based on current literature, posterior combination strategies are most commonly used [26]. The simplest way of combining visual and textual representations is concatenation [30], [31]. However, simple concatenation may bring semantic drift due to the vector space discrepancy of vision and language. Silberer *et al.* represent multimodal data with autoencoders [32]. Encoders are fed with pre-learned visual and text features, and the hidden representations are then used as multimodal embeddings. Collell *et al.* propose to learn a mapping function from text to vision [33]. The outputs of the mapping themselves are used in the multimodal features. However, given an individual item of a modality, there may exist more than one semantically similar items in another modality. Thus, these methods which focus on pairwise similarity cannot effectively identify the mapping relationships of items in different modalities.

To address this limitation, we propose to use adversarial learning. Adversarial learning is demonstrated effective for various applications, like unsupervised domain adaptation to enforce domain-invariant features [11], [34], [35], and regularizing correlation loss between cross-modal items [36]. Adversarial learning provides an alternatively way to generate modality-invariant representations without being restricted by item similarity. To the best of our knowledge, though the adversarial learning approach has been proved successful in many areas [37]–[40], it has not been effectively evaluated in multimodal NER tasks.

C. Bilinear Attention Network

Attention mechanism is widely used in a variety of deep learning tasks [6], [41], [42]. Anderson *et al.* mention that visual objects are much more natural bases for attention [43] and propose a bottom-up attention model for Visual Question Answering task [42]. Object-level features are considered as fine-grained visual features and may be helpful to extract entities which are related to different visual objects in multimodal NER task.

For multimodal NER (MNER) task, a modality attention, which focuses on image, word and character level representation has been proposed in [7]. Their method only takes the attention of textual spans and single images into account. [6] propose an adaptive co-attention model where text and image attentions are captured simultaneously. There are variants of co-attention networks in recent years [44], [45]. However, these co-attention methods use separate attention distributions for each modality, neglecting the interaction between the modalities.

Bilinear attention network is considered good at exploiting the interactions between two groups of input channels [10]. We introduce bilinear attention network in multimodal NER task for capturing the correlations of visual objects and textual entities. We further extend the bilinear attention with a gated module to filter out the irrelevant visual objects.

III. Adversarial Gated Bilinear Attention Network

A. Problem Statement

In this paper, our goal is to predict the named entity label sequence of a given multimedia post. Formally, given an input sentence as $T = (t_1, t_2, \ldots, t_n)$, where t_i denotes the i-th token in the sentence and n is the sentence length, and an input image as V, our task is to predict a label sequence $Y = (y_1, y_2, \ldots, y_n)$ (e.g., in standard BIO2 formats [46]) corresponding to the sentence T and image V, where y_i denotes the label of i-th token in the sentence.

We tackle the multimodal NER task with an adversarial gated bilinear attention network model. Fig. 3. shows the overall architecture of our model. In order to bridge the modality gap between image and text, adversarial learning is adopted to project features from different modalities into a modality-invariant subspace. The adversarial learning consists of two feature projectors (the textual feature projector and the visual feature projector) and one modality classifier. We first project features of the sentence and the image into $\mathcal{G}_T = f_T(T, \theta_T)$ and $\mathcal{G}_V = f_V(V, \theta_V)$, respectively, where f_T is the textual feature projector with parameters θ_T, f_V is the visual feature projector with parameters θ_V, $\mathcal{G}_T$ and $\mathcal{G}_V$ are the projected textual features and projected visual features in the modality-invariant subspace, respectively.

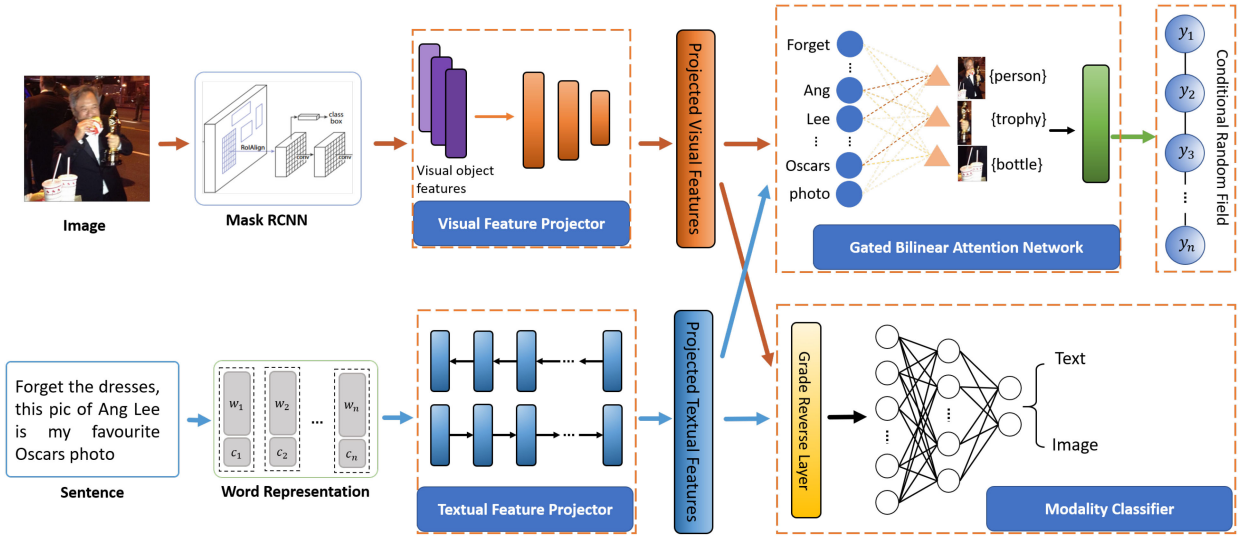


Fig. 3. The overall architecture of our model. Our model utilizes the relevant objects as auxiliary contexts for understanding entities in sentences. Adversarial learning is used to bridge the modality gap between text and visual. A gated bilinear attention network is applied to align entities with related visual objects and fuse multimodal information.

Then, the projected textual and visual features are input to the modality classifier. The feature projectors and modality classifier "play" the min-max game to steer the representation learning. After that, a bilinear attention network (BAN) is introduced to capture the relations of textual entities and visual objects. A gated module is used following the BAN to filter out completely irrelevant visual objects and outputs the multimodal representations. Finally, the conditional random field (CRF) layer takes the multimodal representations as input to predict a NER label sequence.

B. Textual Feature Projector

Similar to the state-of-the-art NER approaches [16], [47], we use both word embeddings and character embeddings to represent the token. Given the input sentence, each token t_i is first embedded in latent space by word embedding:

$$\mathbf{x}_i^w = \mathbf{e}^w(t_i), \tag{1}$$

where $\mathbf{e}^w$ denotes a word embeddings lookup table and $\mathbf{x}_i^w \in \mathbb{R}^{d_w}$. We use pre-trained 200-dimensional GloVe [48] embeddings to initialize it.

Previous works have shown that character representation can improve the NER performance by capturing morphological and semantic information [8], [16]. Similarly to [16], we use Bi-LSTM to extract the character representations which takes a sequence of character of each token as input. The token of input sentence can be seen as a character sequence: $t_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,m}\}$, where $c_{i,j}$ denotes the j-th character for the i-th token in the sentence and m is the token length. We first take the character embedding $\mathbf{e}^c(c_{i,j})$ to represent each character $c_{i,j}$, where $\mathbf{e}^c$ denotes a character embeddings lookup table which is initialized randomly. Then we feed the embedding $\mathbf{x}_{i,j}^c \in \mathbb{R}^{d_{ce}}$ of each character of the token t_i into the bi-directional LSTM to learn hidden states $\overrightarrow{\mathbf{h}}_{i,1}^c, \ldots, \overrightarrow{\mathbf{h}}_{i,m}^c$ and $\overleftarrow{\mathbf{h}}_{i,1}^c, \ldots, \overleftarrow{\mathbf{h}}_{i,m}^c$. The final character representations for t_i is the concatenation of the forward and backward hidden states:

$$\mathbf{x}_i^c = [\overrightarrow{\mathbf{h}}_{i,m}^c; \overleftarrow{\mathbf{h}}_{i,1}^c], \tag{2}$$

where $\mathbf{x}_i^c \in \mathbb{R}^{d_c}$, and d_c is the hidden state dimension of this Bi-LSTM. The token representation $\mathbf{x}_i^t$ is obtained by concatenating $\mathbf{x}_i^w$ and $\mathbf{x}_i^c$:

$$\mathbf{x}_i^t = [\mathbf{x}_i^w; \mathbf{x}_i^c], \tag{3}$$

where $\mathbf{x}_i^t \in \mathbb{R}^{d_w + d_c}$. To capture the contextual information, we send the token representation $\mathbf{x}_i^t$ into another bi-directional LSTM. The forward and backward hidden state of the token t_i can be expressed as follows:

$$\overrightarrow{\mathbf{h}}_i^t = \overrightarrow{\text{LSTM}}(\mathbf{x}_i^t, \overrightarrow{\mathbf{h}}_{i-1}^t), \tag{4}$$

$$\overleftarrow{\mathbf{h}}_i^t = \overleftarrow{\text{LSTM}}(\mathbf{x}_i^t, \overleftarrow{\mathbf{h}}_{i-1}^t). \tag{5}$$

After that, the projected textual features are represented by concatenating the forward and backward hidden state of the token t_i as follow:

$$\mathbf{h}_i^t = [\overrightarrow{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i] \tag{6}$$

where $\mathbf{h}_i^t \in \mathbb{R}^d$ denotes the projected token features, which can be interpreted as representation summarizing the token t_i. The sentence projected features can be denoted as $\mathcal{G}_T = \{\mathbf{h}_1^t, \mathbf{h}_2^t, \ldots, \mathbf{h}_n^t\} \in \mathbb{R}^{d \times n}$, where n is the number of tokens in the sentence. We define θ_T as the parameter set of embedding layers and Bi-LSTM layers.

C. Visual Feature Projector

Image information in social media posts provides auxiliary contexts for understanding entities in sentences. However, the image-level information cannot help to extract entities with different type. With the guidance of object-level information, we

can extract different type of entities corresponding to different objects. For example, one can know that "Ang Lee" is a person name with the guiding visual object "person," and "Oscars" is an award name with the guiding visual object "trophy". Different from previous multimodal named entity recognition works [6], [7], [9], we utilize the object detection model Mask RCNN [49] which is pre-trained on the COCO dataset [50] to recognize the objects in images. We chose the output of last pooling layer of Mask RCNN as the visual object features, which contains the discriminative information describing the semantic of each object. In most cases, only the salient objects are related to the entities mentioned in a sentence. Thus, we consider the top k objects with the highest object classification probabilities, denoting as $\tilde{\mathbf{v}} = \{\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_k\} \in \mathbb{R}^{1024\times k}$, where $\tilde{\mathbf{v}}_i \in \mathbb{R}^{1024}$ denotes the features of the i-th object.

To ensure that feature projector has enough capacity of representations capturing large margins of statistical properties between the image and text modality, we select feed-forward network as the feature projector, denoted as f_V with parameters θ_V. The visual feature projector maps the object features from Mask RCNN into new vector with the same dimensions as the projected textual features:

$$\mathbf{v}_i = f_V(\tilde{\mathbf{v}}_i; \theta_V), \tag{7}$$

where $\mathbf{v}_i \in \mathbb{R}^d$ is the projected visual features of i-th object. Thus, the projected features of object set can be denoted as $\mathcal{G}_V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} \in \mathbb{R}^{d\times k}$.

D. Adversarial Learning on Modality Classifier

After obtaining the projected textual and visual features, we next input these projected features into modality classifier, which guides the process of modality-invariant feature learning. The modality classifier will distinguish whether the input features come from visual feature projector or textual feature projector, which acts as a "discriminator" in GAN (Generative Adversarial Networks) [51]. In our experiments, the modality classifier D is constructed by a 3-layer feed-forward network with parameters θ_D (see Section IV-A2 for implementation details), and a softmax layer is applied to obtain the probability distribution over classes.

To train the modality classifier, the projected textual features are assigned to the label 1, while the projected visual features are assigned to the label 0. Then, the adversarial loss $\mathcal{L}_{adv}$ can be defined as:

$$\begin{aligned}\mathcal{L}_{adv}(\theta_T, \theta_V, \theta_D) = & -\frac{1}{n+k}\sum_{i=1}^{n+k}(\tau_i log(D(\mathbf{o}_i; \theta_D)) \\ & + (1-\tau_i) log(1 - D(\mathbf{o}_i; \theta_D))),\end{aligned} \tag{8}$$

where n is the number of words in the sentence, k is the number of objects in the image, $\mathbf{o}_i \in \{\mathbf{h}_j^t, \mathbf{v}_l\}, j = 1, \ldots, n$ and $l = 1, \ldots, k$ is the token features from the textual feature projector or object features from the visual feature projector. τ_i is the ground-truth modality label of $\mathbf{o}_i$, $D(\mathbf{o}_i; \theta_D)$ is the generated modality probability of the projected feature $\mathbf{o}_i$.

E. Gated Bilinear Attention Network

To better map relations of different visual objects and textual entities, we boost the bilinear attention network with gated module as show in Fig. 4. Bilinear attention network (BAN) has succeeded in multimodal learning task [10], which can exploit the interaction between the modalities by computing the correlation between each pair of the input channels.

Given the sentence projected features $\mathcal{G}_T \in \mathbb{R}^{d\times n}$ obtained by the textual feature projector and the projected features of object set $\mathcal{G}_V \in \mathbb{R}^{d\times k}$ obtained by visual projector, where d is the number of feature dimension of each token or object. As shown in the upper part of Fig. 4, we first reduce both modality feature dimension simultaneously, and then calculate the semantic similarity of each object and token. The attention map $\mathcal{A}$ can be formally defined as:

$$\mathcal{A} = \text{softmax}\left(\left(\left(\mathbb{1}\cdot\mathbf{p}^T\right)\circ\mathcal{G}_T^T\mathbf{W}_t\right)\mathbf{W}_v^T\mathcal{G}_V\right), \tag{9}$$

where $\mathbb{1} \in \mathbb{R}^n$ is a vector of ones, $\mathbf{p} \in \mathbb{R}^{d'}$ is a pooling parameter matrix which takes each element of the hadamard product vector between tokens and objects into account and gains the attention scores. $\mathbf{W}_t \in \mathbb{R}^{d\times d'}$ and $\mathbf{W}_v \in \mathbb{R}^{d\times d'}$ are parameter matrices, $\circ$ is Hadamard product (element-wise multiplication), and remind that $\mathcal{A} \in \mathbb{R}^{n\times k}$. The softmax function is applied element-wisely. Note that each element $\mathcal{A}_{i,j}$ in the bilinear attention map $\mathcal{A}$ denotes the relationship between i-th token in the sentence and the j-th object in the image.

Then, we can calculate the attention based visual object features for each token in sentence as follows:

$$\mathcal{G}_V^{att} = (\mathbf{W}_a\mathcal{G}_V)\mathcal{A}^T. \tag{10}$$

where $\mathbf{W}_a \in \mathbb{R}^{d\times d}$ is a parameter matrix, and $\mathcal{G}_V^{att} \in \mathbb{R}^{d\times n}$.

Consider that not all objects in images are relevant to entities in sentences. For example, as shown in Fig. 3, "bottle" is an irrelevant visual object for recognizing "Ang Lee" as a person name or "Oscars" as an award name. We design a filter gate which can determine if the object is related to the words in sentences. For calculation convenience, we leverage a single layer perceptron with activation function $tanh$ to project the attention based visual features and textual features into lower dimensions. We next compute a multimodal gate with sigmoid function as shown in the lower part of Fig. 4. The Gated Filter is defined as:

$$\mathbf{h}_v = tanh(\mathbf{W}_{gv}\mathcal{G}_V^{att}), \tag{11}$$

$$\mathbf{h}_t = tanh(\mathbf{W}_{gt}\mathcal{G}_T), \tag{12}$$

$$\mathbf{g} = \sigma(\mathbf{W}_g[\mathbf{h}_v; \mathbf{h}_t]), \tag{13}$$

where $\mathbf{W}_{gv} \in \mathbb{R}^{d_g\times d}$, $\mathbf{W}_{gt} \in \mathbb{R}^{d_g\times d}$ and $\mathbf{W}_g \in \mathbb{R}^{d\times 2d_g}$ are trainable parameters, σ is the logistic sigmoid activation which be used as gated function. $\mathbf{h}_v \in \mathbb{R}^{d_g\times n}$, $\mathbf{h}_t \in \mathbb{R}^{d_g\times n}$ and $\mathbf{g} \in \mathbb{R}^{d\times n}$.

After that, we apply the multimodal gate to filter useless visual object information from the attention based visual feature. Finally, we add the gated attention based visual features and projected textual features to generate the multimodal features:

$$\mathcal{G}_M = \mathbf{g}\circ\mathcal{G}_V^{att} + \mathcal{G}_T. \tag{14}$$

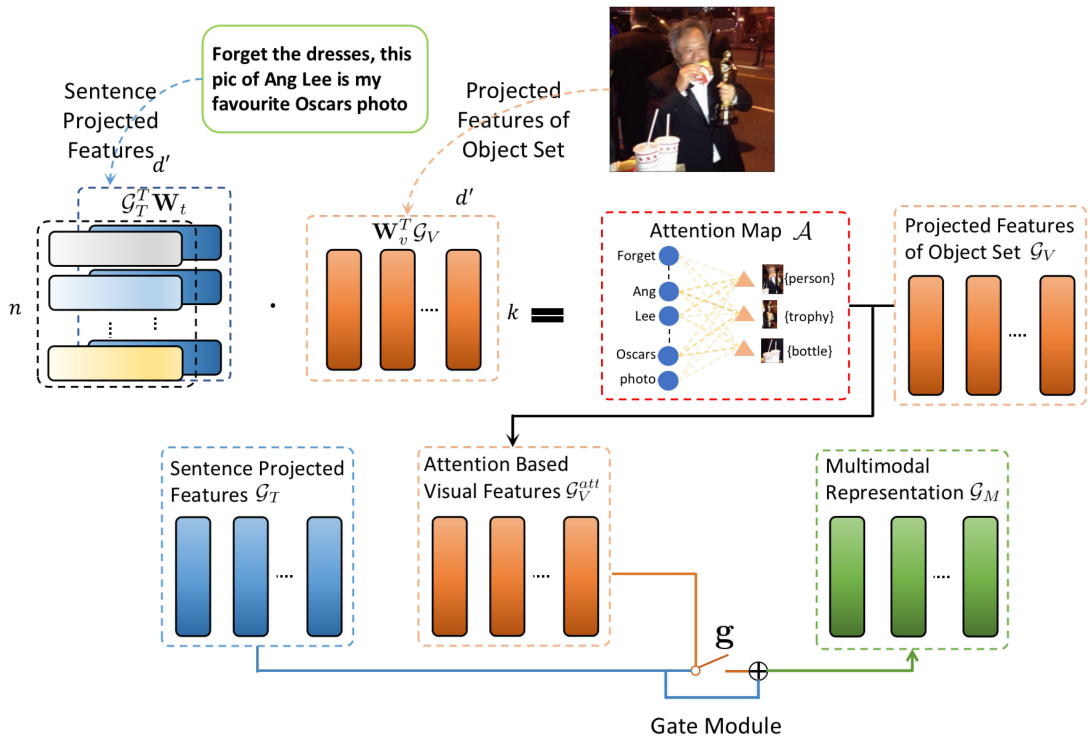


Fig. 4. Overview of gated bilinear attention network. The two modality features, sentence projected features and projected features of object set, are used to get bilinear attention maps. A gated module is designed to filter out the irrelevant visual objects, which results in the multimodal representation. The $\oplus$ sign in the Gate Module indicates add operation.

where $\mathcal{G}_M \in \mathbb{R}^{d \times n}$ is the multimodal features of the sentence. We define θ_{GBAN} as the parameter set of all parameter matrices $\mathbf{W}$ and $\mathbf{p}$ in the Gated Bilinear Attention Network.

The gated module is designed to filter out irrelevant visual objects by considering the relations between objects and entities, as shown in Equation (11–14). However, to preserve the original textual information, which we think is more important in identifying entities, we add the filtered object features $\mathbf{g} \circ \mathcal{G}_V^{att}$ with textual features $\mathcal{G}_T$ to get the final multimodal features.

F. Output Layer

In the named entity recognition task, the correlations between labels in neighborhoods are beneficial to predict correctly the chain of labels. For example, I-PER cannot follow I-OTHER in BIO2 annotation. Therefore, instead of modeling tagging decisions independently, we use Conditional Random Fields (CRF) to model the label dependence, which has been shown to improve named entity recognition task. For an input sentence:

$$T = (t_1, t_2, \ldots, t_n), \tag{15}$$

in order to get the score matrix, we apply the single feed forward layer to the multimodal features obtained by the gated bilinear attention network:

$$\mathbf{S} = \mathcal{G}_M^T \mathbf{W}_s, \tag{16}$$

where $\mathbf{W}_s \in \mathbb{R}^{d \times d_l}$ is a parameter matrix, $\mathbf{S}$ has size $n \times d_l$, d_l is the number of distinct tags, and $\mathbf{S}_{i,j}$ is the score of the j-th tag of the i-th token in a sentence. For a sequence of predictions:

$$Y = (y_1, y_2, \ldots, y_n), \tag{17}$$

the score of Y can be defined as follows, which takes the model prediction and the dependencies among contexts into consideration:

$$\Phi(T, Y) = \sum_{i=1}^{n-1} \mathbf{Q}_{y_i, y_{i+1}} + \sum_{i=1}^{n} \mathbf{S}_{i, y_i}, \tag{18}$$

where $\mathbf{Q} \in \mathbb{R}^{d_l \times d_l}$ is a parameter matrix of transition scores, and $\mathbf{Q}_{i,j}$ represents the score of a transition from the tag i to tag j. Then, a softmax is applied to all possible tag sequences to yield a probability for the sequence Y:

$$p(Y|T) = \frac{e^{\Phi(T,Y)}}{\sum_{\tilde{Y} \in Y_T} e^{\Phi(T,\tilde{Y})}}, \tag{19}$$

where Y_T is the set of all possible tag sequences for a sentence T. During training phase, we maximize the log-probability of the correct label sequence. So the CRF loss can formally defined as follows:

$$\begin{aligned} \mathcal{L}_{crf}(\theta_V, \theta_T, \theta_{GBAN}, \theta_{CRF}) &= -log(p(Y|T)) \\ &= log\left(\sum_{\tilde{Y} \in Y_T} e^{\Phi(T,\tilde{Y})}\right) - \Phi(T, Y), \end{aligned} \tag{20}$$

where θ_{CRF} denotes the parameter matrices $\mathbf{W}_s$ and $\mathbf{Q}$ of CRF layer. We optimize our network to produce a valid sequence of

Algorithm 1: Pseudocode of Optimizing Our Proposed AGBAN Model.

Initialization: Images for current batch $\mathcal{V} = \{v_0, v_1, \ldots, v_N\}$, sentences for current batch $\mathcal{T} = \{T_0, T_1, \ldots, T_N\}$, corresponding labels for current batch $\mathcal{Y} = \{Y_0, Y_1, \ldots, Y_N\}$, γ samples in a minibatch.

update until convergence:

for $z - steps$ **do**
  update parameters $\theta_V, \theta_T, \theta_{GBAN}$ and θ_{CRF} by **descending** their stochastic gradients:
  $\theta_V \leftarrow \theta_V - \mu \cdot \nabla_{\theta_V} \frac{1}{\gamma}(\mathcal{L}_{crf} - \mathcal{L}_{adv})$
  $\theta_T \leftarrow \theta_T - \mu \cdot \nabla_{\theta_T} \frac{1}{\gamma}(\mathcal{L}_{crf} - \mathcal{L}_{adv})$
  $\theta_{GBAN} \leftarrow \theta_{GBAN} - \mu \cdot \nabla_{\theta_{GBAN}} \frac{1}{\gamma}(\mathcal{L}_{crf} - \mathcal{L}_{adv})$
  $\theta_{CRF} \leftarrow \theta_{CRF} - \mu \cdot \nabla_{\theta_{CRF}} \frac{1}{\gamma}(\mathcal{L}_{crf} - \mathcal{L}_{adv})$
end for
update parameters of modality classifier by **ascending** its stochastic gradients through Gradient Reversal Layer:
$\theta_D \leftarrow \theta_D + \mu \cdot \nabla_{\theta_D} \frac{1}{\gamma}(\mathcal{L}_{crf} - \mathcal{L}_{adv})$
returnlearned parameters in the AGBAN model.

output labels by minimizing the CRF loss $\mathcal{L}_{crf}$. During decoding, the label sequence Y^* with the highest conditional probability is selected as output label sequence:

$$Y^* = \underset{\tilde{Y} \in Y_T}{\operatorname{argmax}} p(\tilde{Y}|T). \tag{21}$$

G. Optimization for CRF and Adversarial Learning

As mentioned above, at training, we maximize the adversarial loss $\mathcal{L}_{adv}$ to optimize the parameters $(\theta_V, \theta_T, \theta_D)$, which can make the two modality distributions similar. At the same time, we minimize the CRF loss $\mathcal{L}_{CRF}$ to optimize the parameters $(\theta_V, \theta_T, \theta_{GBAN}, \theta_{CRF})$, which can keep the task-specific information needed to complete named entity recognition. Since the goal of these two optimizations sub-process is opposite, the total optimization process can be formally described as a min-max game:

$$\left(\hat{\theta}_V, \hat{\theta}_T, \hat{\theta}_{GBAN}, \hat{\theta}_{CRF}\right) = \underset{\theta_V, \theta_T, \theta_{GBAN}, \theta_{CRF}}{\arg\min} \left(\mathcal{L}_{crf}(\theta_V, \theta_T, \theta_{GBAN}, \theta_{CRF}) - \mathcal{L}_{adv}(\theta_V, \theta_T, \hat{\theta}_D)\right), \tag{22}$$

$$\hat{\theta}_D = \underset{\theta_D}{\arg\max}(\mathcal{L}_{crf}(\hat{\theta}_V, \hat{\theta}_T, \hat{\theta}_{GBAN}, \hat{\theta}_{CRF}) - \mathcal{L}_{adv}(\hat{\theta}_V, \hat{\theta}_T, \theta_D)). \tag{23}$$

Following the work of Ganin *et al.* [11], we apply the Gradient Reversal Layer (GRL) for our model to implement this min-max game as shown in Fig. 3. There is no other parameters in GRL except the meta-parameter λ. The GRL acts as an identity transform during the forward propagation, while multiplying the gradient by $-\lambda$ and passing the processed gradient to preceding layers when back-propagating. Mathematically, the gradient reversal layer $R_\lambda(x)$ can be formulated as:

$$R_\lambda(x) = x, \tag{24}$$

$$\frac{dR_\lambda}{dx} = -\lambda \mathbf{I}, \tag{25}$$

where $\mathbf{I}$ is an identity matrix. As shown in the Algorithm 1, the min-max optimization game can be performed simultaneously with the GRL layer.

TABLE I
STATISTICS OF THE TWITTER DATASET

	Train	Dev	Test	Total
Person	2217	552	1816	4583
Location	2091	522	1697	4308
Organization	928	247	839	2012
Misc	940	225	726	1881
Total Entity	6176	1546	5078	12784

TABLE II
HYPERPARAMETERS

Parameter	Value	Parameter	Value
word emb dim d_w	200	selected object num k	3
char emb dim d_{ce}	30	GBAN param dim d'	200
char hidden dim d_c	50	Gate param dim d_g	200
word hidden dim d	200	CRF param dim d_l	9
dropout rate	0.5	L_2 regularization λ	1e-8
batch size	10	learning rate decay	0.05
learning rate	0.005	$z - steps$	8

IV. EXPERIMENT

A. Experimental Setup

1) Dataset: We evaluate our model on a multimodal social media dataset [6] from Twitter. It contains 8257 tweets posted by 2116 user. The dataset contains four different types of entities: **Person**, **Location**, **Organization**, **Misc**. We leverage the standard BIO2 tagging scheme as most previous NER works [6], [16] in which non-entity is tagged with label O. The total number of entities is 12784. Following the same settings as zhang *et al.* [6], we split the dataset into training set, development set and testing set, which contain 4000, 1000 and 3257 tweets, respectively. The statistics of each type of named entities in training, development and test sets are shown in Table I.

2) Implementation Details: Table II shows the hyperparameters used in our experiments, which mostly follow yang *et al.* [47]. GloVe 200-dimension [48] is used to initialize word embeddings. The out of vocabulary (OOV) words are initialized by randomly sampling from a uniform distribution of $[-\sqrt{\frac{3}{d_w}}, \sqrt{\frac{3}{d_w}}]$, where d_w is the dimension of the word embeddings [52]. The weights of the character embedding layer are randomly initialized from a uniform distribution of $[-\sqrt{\frac{3}{d_{ce}}}, \sqrt{\frac{3}{d_{ce}}}]$, where d_{ce} is the dimension of character embeddings and we set $d_{ce} = 30$. The word and char embeddings are fine-tuned during training. The hidden dimensions of char-level Bi-LSTM and word-level Bi-LSTM are set to 50 and 200, respectively. We

choose the optimal number k of selected object based on the validation results. The dimensions of parameter matrices for GBAN and Gate module are 200. The size of CRF transition parameter matrix is 9 which corresponds to the label counts (e.g., for the label PERSON, we tag the beginning and the other words with B-PER and I-PER).

We implement our model on PyTorch framework.[1] To avoid overfitting, we apply dropout [54] to both word and char embeddings with a rate of 0.5. The mini-batch stochastic gradient descent (SGD) with a decayed learning rate is used to update parameters. We set the batch size to 10, $k - steps$ to 8 and the learning rate to 0.005. The learning rate decay is 0.05.

3) Evaluation Metric: Standard precision, recall and F1-score are used as the evaluation metrics for our experiments. We select the optimal model base on the performance on development datasets, and report the performance of selected model on the test dataset.

4) Baseline Methods: To validate the effectiveness of our model, we compare our model with several baseline models. Our experiments mainly concern two groups of models: the state-of-the-art models and the variants of our model.

Previous State-of-the-art Models: we compare our model with following existing state-of-the art models.

- **Stanford NER:** The Stanford NER is a tool that is widely used to handle named entity recognition task. It was proposed by finkle *et al.* [53].
- **T-NER:** T-NER [5] is a model that focuses on named entity recognition in tweets domain. A set of widely-used effective features are used in T-NER, such as, dictionary, contextual and orthographic features.
- **CNN+BiLSTM+CRF:** This is an end-to-end NER system proposed by ma *et al.* [15]. It introduces the character-level representation to enhance the feature representation for each word and achieved the best result on the CoNLL 2003 test set as the author reported.
- **MNER-MA:** The multimodal NER model proposed by moon *et al.* [7] incorporates visual information with a modality attention module. Since they did not provide the data and code used in their paper, we reimplement their model following the same settings and validate on the Twitter dataset.
- **VAM:** The visual attention model [9] is another neural model for multimodal NER task. This model is composed by a BiLSTM-CRF model and a designed visual attention model. Also, we reimplement their model following the same settings of their paper since they did not provide the code and data.
- **AdapCoAtt Model:** The adaptive co-attention network proposed by zhang *et al.* [6] is a multimodal model for named entity recognition in tweets, which combines the visual information and textual information.
- **BERT-NER:** We compare our model with contextual language models to show the effectiveness of combining object-level features. We use the BERT BASE model [17] and fine-tuning in the Twitter dataset.

[1]https://pytorch.org/

Variants of Our Model: We set the ablation experiments to illustrate the contribution of each component in our model. For fair comparison, we use the same parameter settings for each model.

- **Object-Concat (Text + Object + Concatenate):** This model is a baseline of our model without Bilinear Attention Network module, Gated module and Adversarial Learning module. Instead of using Gated Bilinear Attention, we simply concatenate the object features with textual representations. Specifically, we simply concatenate all the object features as the global visual features and use the feed forward layer to fuse different visual object features. Then, the new global visual features are added to each word in sentences for the multimodal features.
- **Object-BAN (Text + Object + BAN):** This is a variant of our model based on **Object-Concat** model, while we add the Bilinear Attention Network to capture the correlations between words and visual objects. And we directly add the attention based visual features to the word features for each word without the Gated module.
- **Object-GBAN (Text + Object + Gated + BAN):** This is another variant of our model based on **Object-BAN**, with the Gated module to filter out the useless visual object features. For all our models without adversarial training, the feature projector is trained by deleting the adversarial loss from Equation (22).
- **Object-AGBAN (Text + Object + Adversarial Learning + Gated + BAN):** This is the complete model. It utilizes the Adversarial Learning to bridge the modality gap between word features and object features. The Bilinear Attention Network is used to find the mapping relations between words and objects, and Gated module in Bilinear Attention Network is designed to filter out the irrelevant visual object features. Finally, the multimodal features are input to CRF layer for inference.

B. Result and Discussion

1) Comparison With Existing Models: We first compare our model with 7 state-of-the-art models on Twitter dataset. Table III shows the testing results of compared models and our models. Our Object-AGBAN model significantly outperforms the compared models in F1 value. CNN+BiLSTM+CRF [15] is a textual baseline of our models without the visual information. The state-of-the-art multimodal NER methods: MNER-MA [7], VAM [9] and AdapCoAtt Model [6], outperform the textual baseline by considering the image-level features. However, we show that when incorporating the object-level features, the model performance is improved from 70.69% to 73.25%. We also compare our model with pre-trained contextual language model BERT-NER [17]. Our variant model Object-Concat gains comparable results against BERT-NER, and we show that with the Gated Bilinear Attention module and Adversarial Learning, our model can outperform the BERT-NER in Precision and F1 values.

2) Ablation Study: The second part of Table III shows the performances of our variant models. The results demonstrate that

TABLE III
RESULTS ON THE TWITTER TEST SET. THE FIRST PART IS RESULTS OF SEVERAL STATE-OF-THE-ART MODELS. THE SECOND PART IS PERFORMANCES OF OUR MODELS. RESULTS ON ROWS WHERE THE MODEL NAME IS MARKED WITH A ‡ SYMBOL ARE REPORTED AS PUBLISHED, ALL OTHER NUMBERS HAVE BEEN COMPUTED BY US. $*$ INDICATES THE DIFFERENCE AGAINST THE F1 OF OUR BASELINE VARIANT (OBJECT-CONCAT) IS STATISTICALLY SIGNIFICANT BY ONE-TAILED PAIRED t-TEST WITH $p < 0.01$

Model	Precision	Recall	F1
Stanford NER [53]‡	60.98	62.00	61.48
T-NER [5]‡	69.54	68.65	69.09
CNN+BiLSTM+CRF [15]‡	66.24	68.09	67.15
MNER-MA [7]	72.33	63.51	67.63
VAM [9]	69.09	65.79	67.40
AdapCoAtt Model [6]‡	72.75	68.74	70.69
BERT-NER [17]	70.65	**73.29**	71.87
Object-Concat (Text + Object + Concatenate)	74.09	69.73	71.85
Object-BAN (Text + Object + BAN)	74.72	70.50	72.55*
Object-GBAN (Text + Object + Gated + BAN)	**75.42**	70.44	72.84*
Object-AGBAN (Text + Object + Adversarial Learning + Gated + BAN)	74.13	**72.39**	**73.25***

all the components of our model play a critical role in improving NER performance.

Object-Concat is a baseline of our model without Bilinear Attention Network, Gated module and Adversarial Learning. The F1-score of this model is 71.85%, which is higher than all compared state-of-the-art models (comparable to the BERT-NER model). That demonstrates the effectiveness of using visual object features.

Object-BAN is an improved model against *Object-Concat*, which leverages the Bilinear Attention Network module to capture the relevance between words and objects instead of simple concatenation. By adding the Bilinear Attention Network module, the F1-score is improved from 71.85% to 72.55%, revealing the effectiveness of Bilinear Attention Network module. Particularly, we can find that *Object-BAN* model gains a 74.72% precision score, which is higher than the *Object-Concat* model. A possible reason is that Bilinear Attention Network can exploit the interactions between words in sentences and visual objects in images and better capture the relations of textual entities and visual objects. The fine-grained relevant visual object features help model to extract named entities correctly.

Object-GBAN is based on *Object-BAN*, but a Gated module is added to filter out completely irrelevant visual objects information. We can find that gated module gains a higher precision score, and finally improves F1-score from 72.55% to 72.84% compared to *Object-BAN*.

Object-AGBAN is the complete model composed of *Object-GBAN* and adversarial learning. It uses adversarial learning to bridge the modality gap between visual and textual features. By mapping two modality features to a common subspace, *Object-AGBAN* achieves the highest Recall 72.39% and F1-score 73.25%.

3) Performance on Categories: Table IV shows our results on four categories. Compared to two state-of-the art models [5], [6], the Object-AGBAN achieves the best F1-score on all four entity categories. Specifically, our model achieves more improvements on ORG and MISC categories. A possible explanation is that ORG and MISC entities require more fine-grained visual objects information as auxiliary context to be recognized. For example, we can recognize the ORG entity "Chicago Blackhawks" is a hockey team name with visual object "ball" and "person". And the MISC entity "Paulus" can be recognized as a dog's name, with the help of the visual object "dog" in the image.

TABLE IV
OUR RESULTS ON FOUR CATEGORIES COMPARED TO T-NER AND ADAPCOACTT MODEL ON THE TWITTER TEST SET

Category	Our Model (Object-AGBAN)			T-NER	Adap.
	Prec.	Recall	F1	F1	F1
PER	82.27	87.39	**84.75**	83.64	81.98
LOC	71.65	81.26	**79.41**	76.18	78.95
ORG	63.70	53.75	**58.31**	50.26	53.07
MISC	47.44	35.67	**40.72**	34.56	34.02
Overall	74.13	72.39	**73.25**	69.09	70.69

TABLE V
THE RESULT (F1 VALUE) OF OUR MODEL ON THE TWITTER TEST SET WITH AND WITHOUT DROPOUT

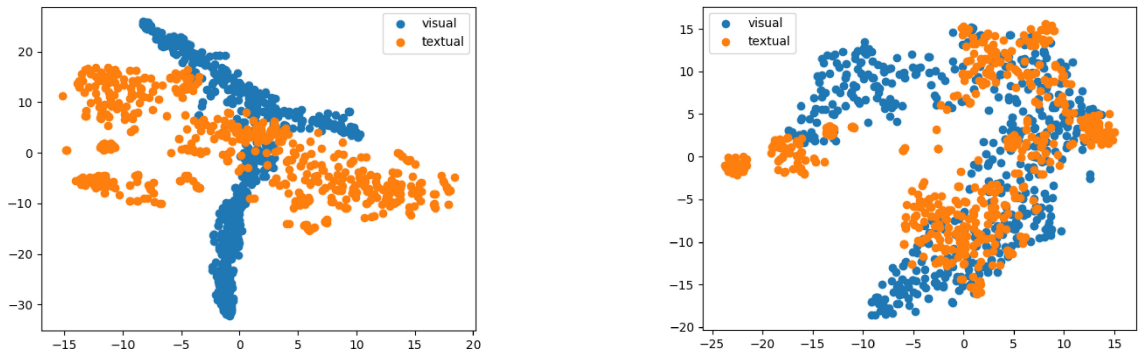
	Overall	PER	LOC	ORG	MISC
No	67.26	79.22	74.23	51.29	28.38
Yes	**73.25**	**84.75**	**79.41**	**58.31**	**40.72**

TABLE VI
RESULTS OF OUR PROPOSED MODEL ON DIFFERENT OBJECT NUMBER

Ojbect Num.	Overall	PER.	LOC.	ORG.	MISC
Top-1	71.01	82.63	77.45	55.13	35.16
Top-2	72.67	83.81	78.92	57.03	38.86
Top-3	**73.25**	84.75	79.41	58.31	40.72
Top-4	72.86	85.10	79.20	57.18	38.60
Top-5	72.81	84.49	79.28	56.26	39.02

4) Parameter Sensitivity: In this section, we further discuss the performance of our model on different settings of the parameters. Specifically, we check the sensitivity of the number of objects k and the impact of dropout.

To illustrate the contribution of dropout, we remove all dropout layers in our model, and keep other hyper-parameters consistent with our model. Table V shows those results of our models on Twitter test set. The performance on each category



(a) Feature visualization without adversarial learning

(b) Feature visualization with adversarial learning

Fig. 5. The t-SNE visualization of test data in Twitter dataset. The orange color points represent textual features and the blue color points represent the visual features.

(a) (b)

Fig. 6. Visualization of the bilinear attention maps for two samples from Twitter test set. In each group, the image is shown in the left, the visualization of the bilinear attention map is shown in the right. The horizontal axis shows sentence and the vertical axis denotes selected object. the color in each grid cell shows the relevance between relative words and objects.

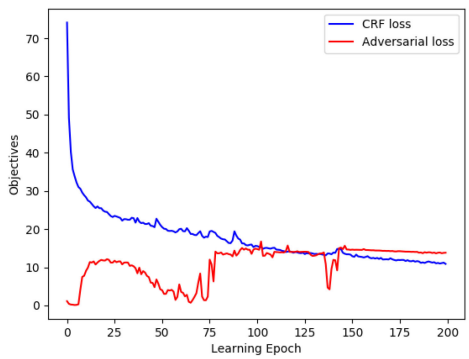


Fig. 7. Development of the adversarial loss and CRF loss of our model on the Twitter dataset during the training process

and overall is improved by adding dropout. The possible reason is that dropout can avoid overfitting. Our model achieved the state-of-the-art performance with dropout.

As mentioned in Section III, we chose the top k detected objects according to the detection possibility of each object in an image. Table VI shows the results of our model with different number of selected objects. As can be noted, the F1-score firstly increases and then decreases with the changes of the selected objects, and we achieve the best F1-score when setting the number of objects to 3. When we only select the most salient object, the performance declines by 2.24%. The reason is that not only salient object is important to recognize different named entities in sentences, small object also can be useful. For example, in Fig. 3., the "trophy" object in image is small, however, it is useful to recognize the "Oscar" as an award name. Meanwhile, too many visual objects may bring noise into the prediction of entities. However, even if given 5 visual objects, the performance is only decreased by 0.44%. The main reason is that the Gated Bilinear Attention Network can filter out irrelevant visual object information.

5) Effect of Adversarial Learning: In our model, we apply adversarial learning to bridge the modality gap between image and text. During training phase, we jointly optimize the CRF loss and adversarial loss in the objective function by using GRF layer. To explore the effect of adversarial learning in our model, we visualize the adversarial loss and the CRF loss from epoch 1 to 200 in Fig. 7. As shown in the Fig. 7, the CRF loss firstly decreases and then converges smoothly. While the adversarial loss firstly increases during 0 to 25 epochs, which illustrates that the feature projector is working to bridge the modality gap. In this case, modality classifier is unable to distinguish exactly which modality the features come from. Then the adversarial loss decreases during 25 to 75 epochs, which means that the modality classifier is getting stronger to better guide the process of feature project. And then the adversarial loss fluctuates and converges until 150 epochs. This denotes the "adversarial learning" between the modality classifier and the feature projector. Finally the adversarial loss keeps stable. These results are in line with our expectation that the modality classifier in our model acts as a discriminator for guiding the direction of forming a common modality subspace. If the value of adversarial loss would explode, modality classifier would be too weak to

(a) Our model vs. CNN+BiLSTM+CRF

(b) Our model vs. AdapCoAtt model

Fig. 8. Four examples for illustrating the effectiveness of objects information in images. The left and right examples are the results of our model comparing to CNN+BiLSTM+CRF [15] and AdapCoAtt Model [6], respectively. In each group, the image is shown in the left, while the sentence with ground-truth, selected objects and predicted entities are showed in the right. Our model recognizes named entities precisely when fine-grained visual objects information is used as auxiliary context.

guide the feature projector to project two modality features into a common space. Conversely, if the adversarial loss decreased to zero, modality classifier would win the minmax game, which means that the feature projector fails to project two modality features into a common space.

To further evaluate the effectiveness of Adversarial Learning, we use t-SNE tool to visualize the distribution of the projected features from our trained model on Twitter dataset. Specifically, we first randomly select 500 tokens or objects from test data in Twitter dataset for each modality. Then, the trained model is used to get the projected features. For training t-SNE, we set the perplexity to 60, and train the t-SNE for 500 iterations. Fig. 5(a) and (b) are the t-SNE visualizations of projected features without and with adversarial learning, respectively. Fig. 5(b) illustrates that adversarial learning can bridge the modality gap and align distributions of different modality features.

6) Visualization of BAN: To validate that our model is able to find the correlations between words in the sentence and objects in the image. We select two samples from Twitter test set, and visualize the Gated Bilinear Attention Network in Fig. 6. In the left group, the sentence is "Chicken Alfredo made for me by the one and only @masonvogel," and the objects selected from image are "cake," "dining table" and "cup". As shown in Fig. 6(a), our bilinear attention network identifies that "Chicken Alfredo" is most relevant to the extracted objects, and the visual object information can help to recognize the "Chicken Alfredo" as a MISC entity. In the right group, the sentence is "RT @KerriFar:RT@mscator: Kolo loves the sun and is so pretty, too. # grilling," and the selected object is "dog". Our bilinear attention map also finds that "dog" is most relevant to "Kolo," which helps model to identify "Kolo" as a dog's name, instead of a name of a person.

7) Case Study: Visual Object information is very important for recognizing named entities in social media posts. We take four examples in Twitter test set for illustrating the effectiveness of our proposed model. Fig. 8(a) shows the predicted results of our model and the CNN+BiLSTM+CRF model. As shown in the first row of Fig. 8(a), even though the CNN+BiLSTM+CRF can locate the named entity "Mickey," it incorrectly classifies its type to "PER". The reason is that the sentence is short and lack of context. With the guiding visual object "cat" in the image, our model can correctly predict "Mickey" as MISC, which is a name of a cat. In the second row of Fig. 8(a), CNN+BiLSTM+CRF incorrectly recognizes the "Plymouth Barracuda" as LOC named entity, and our model correctly predicts the "Plymouth Barracuda" as the car name, i.e., MISC entity, with the guiding visual object "car" in the image.

In the Fig. 8(b), we compare our model with AdapCoAtt Model [6] which utilizes the image-level features as visual representations. As shown in the first row of Fig. 8(b), our model can extract the PER entity "Jesus" and PER entity "Kanye" correctly, while AdapCoAtt Model can only extract the PER entity "Kanye". A possible explanation is that our model detects that the two people in the image are relative to "Jesus" and "Kanye". However, the image-level features used in AdapCoAtt Model can only provide visual information about one person, which resulted in missing another PER entity "Jesus". The sample shown in the second row of Fig. 8(b) also illustrates the advantages of using fine-grained object information. The model identify the "Mike Douglas Show" as a name of a perform, i.e. MISC entity, by using the guidance of visual objects "Person," "HandBag" and "Tie" as auxiliary contexts. However, the image-level AdapCoAtt Model incorrectly extract "Mike Douglas Show" as a name of a person since "person" is the main information of this image.

Fig. 9. shows some failed examples of our proposed Object-AGBAN model. NER performance mostly benefits when the words of sentences are well-aligned with visual objects in images, however, it is not always the case in social media. The examples in Fig. 9 indicates that when sentences lack contexts

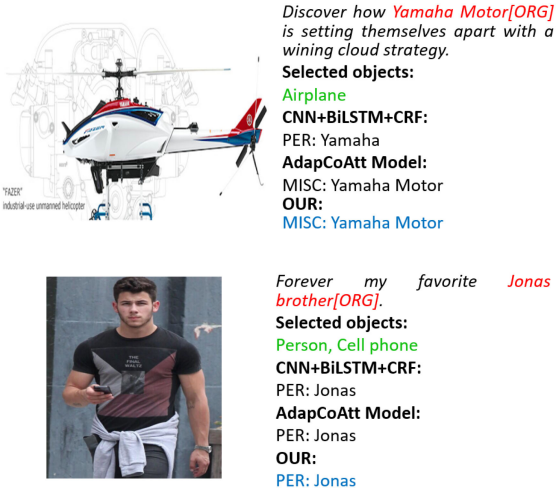


Fig. 9. Error analysis of our proposed Object-AGBAN model. In each group, the image is shown in the left, while the sentence with ground-truth, selected objects and predicted entities are showed in the right. Our model and the compared methods extract the entities incorrectly when visual objects cannot reveal the label semantics of entities.

and also, the extracted objects cannot reveal the label semantics of entities, our model makes the error predictions. In the first row, "Yamaha Motor" is a name of a company. However, since we extract the object "Airplane" in the image, our model incorrectly identifies the entity as a name of a MISC entity. The Text-based CNN+BiLSTM+CRF model extracts "Yamaha" as a person name without the textual contexts. The image-level AdapCoAtt model gets the same results as ours with the identification of an airplane.

The same thing happens in the example in the second row when "Jonas brother" is another company name. Our model extracts the "Jonas" as a person name with the guiding of selected objects "Person" and "Cell phone," the same as the AdapCoAtt model and the CNN+BiLSTM+CRF model.

V. Conclusion

In this paper, we present the adversarial gated bilinear attention network (AGBAN) for multimodal named entity recognition in social media posts. Adversarial learning framework can build the common subspace for different modalities and generate modality-invariant representations bridging vision and language. We extend the bilinear attention network (BAN) with gated mechanism. The BAN exploits bilinear interactions between two groups of input channels and the gated module can filter out irrelevant visual information.

The object-level visual features contribute significantly to the final NER results. Experimental results demonstrate that the gated bilinear attention network can capture the correlations of visual objects and textual entities which helps to extract entities precisely.

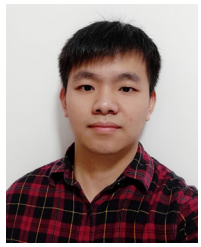
For future work, we consider to combine knowledge-based methods in our multimodal representation for a more robust and effective NER model.

References

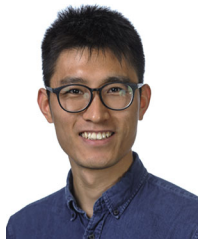
[1] A. Ritter, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1104–1112.
[2] A. Ritter, E. Wright, W. Casey, and T. Mitchell, "Weakly supervised extraction of computer security events from twitter," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 896–905.
[3] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in *Proc. 5th Int. AAAI Conf. Weblogs Soc. Media*, 2011, pp. 265–272.
[4] C. Li, A. Sun, J. Weng, and Q. He, "Tweet segmentation and its application to named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 2, pp. 558–570, Feb. 2015.
[5] A. Ritter *et al.*, "Named entity recognition in tweets: An experimental study," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1524–1534.
[6] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5674–5681.
[7] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., Vol. 1 (Long Papers)*, 2018, pp. 852–860.
[8] Y. Lin, S. Yang, V. Stoyanov, and H. Ji, "A multi-lingual multi-task architecture for low-resource sequence labeling," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, vol. 1, pp. 799–809.
[9] D. Lu, L. Neves, V. Carvalho, N. Zhang, and H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, vol. 1, pp. 1990–1999.
[10] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 1564–1574.
[11] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[12] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.
[13] N. Gupta, S. Singh, and D. Roth, "Entity linking via joint encoding of types, descriptions, and context," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2681–2690.
[14] S. Zheng *et al.*, "Joint extraction of entities and relations based on a novel tagging scheme," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, vol. 1, pp. 1227–1236.
[15] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2016, pp. 1064–1074.
[16] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2016, pp. 260–270.
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., Vol. 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
[18] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1638–1649.
[19] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919–931, Apr. 2013.
[20] I. Varga *et al.*, "Aid is out there: Looking for help from tweets during a large scale disaster," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, vol. 1, pp. 1619–1629.
[21] R. P. Khandpur *et al.*, "Crowdsourcing cybersecurity: Cyber attack detection using social media," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 1049–1057.
[22] S. Phuvipadawat and T. Murata, "Breaking news detection and tracking in twitter," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2010, vol. 3, pp. 120–123.
[23] K. Gimpel *et al.*, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Lang. Technol.*, 2011, pp. 42–47.
[24] T. Baldwin *et al.*, "Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition," in *Proc. Workshop Noisy User-Generated Text*, 2015, pp. 126–135.

[25] G. Aguilar, S. Maharjan, A. P. López-Monroy, and T. Solorio, "A multi-task approach for named entity recognition in social media data," in *Proc. 3rd Workshop Noisy User-generated Text*, 2017, pp. 148–153.
[26] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[27] D. Wang and K. Mao, "Learning semantic text features for web text-aided image classification," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 2985–2996, Dec. 2019.
[28] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 997–1007, Apr. 2018.
[29] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, Apr. 2019.
[30] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 36–45.
[31] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *J. Artif. Intell. Res.*, vol. 49, pp. 1–47, 2014.
[32] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, vol. 1, pp. 721–732.
[33] G. Collell, T. Zhang, and M.-F. Moens, "Imagined visual representations as multimodal embeddings," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4378–4384.
[34] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 557–570, 2018.
[35] Y. Zhang, R. Barzilay, and T. Jaakkola, "Aspect-augmented adversarial networks for domain adaptation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 515–528, 2017.
[36] L. He *et al.*, "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2017, pp. 1153–1158.
[37] J. Li *et al.*, "Adversarial learning for neural dialogue generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2157–2169.
[38] X. Chen *et al.*, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.
[39] J. Wang *et al.*, "Irgan: A minimax game for unifying generative and discriminative information retrieval models," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 515–524.
[40] Y. Tang and X. Wu, "Salient object detection using cascaded convolutional neural networks and adversarial learning," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2237–2247, Sep. 2019.
[41] S. Sukhbaatar *et al.*, "End-to-end memory networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.
[42] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.
[43] R. Egly, J. Driver, and R. D. Rafal, "Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects," *J. Exp. Psychol.: General*, vol. 123, no. 2, pp. 161–177, 1994.
[44] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 289–297.
[45] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 299–307.
[46] E. F. Sang and J. Veenstra, "Representing text chunks," in *Proc. 9th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 1999, pp. 173–179.
[47] J. Yang and Y. Zhang, "Ncrf++: An open-source neural sequence labeling toolkit," in *Proc. ACL 2018, Syst. Demonstrations*, 2018, pp. 74–79.
[48] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[49] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[50] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[51] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[53] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, 2005, pp. 363–370.
[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

Changmeng Zheng received the B.S. degree from the School of Software Engineering, South China University of Technology, Guangdong, China, where he is currently working toward the master's degree. His research interests include multimedia information retrieval and social media analytics.



Zhiwei Wu received the B.S. degree from the School of Software Engineering, South China University of Technology, Guangdong, China, where he is currently working toward the master's degree. His research interests include multimedia and knowledge graph.



Tao Wang is a research associate in the Department of Biostatistics and Health Informatics, King's College London. He received his Ph.D. degree in economics from the University of Southampton, jointly trained with The Alan Turing Institute, U.K. His research interests include text mining, social networks and health informatics.

Yi Cai (Member, IEEE) received the Ph.D. degree in computer science from the Chinese University of Hong Kong. He is currently a Professor with the South China University of Technology (SCUT). His research interests include recommendation system, personalized search, semantic web and data mining. His research works are published on many conferences and journals, such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), *Neural Networks, Knowledge-based Systems, EAAI and Neurocomputing*, as well as AAAI, COLING, CIKM, AAMAS, DASFAA and other international conferences about perspective mining, cognitive modeling, information retrieval and semantic analysis. He also received the National Science and Technology Academic Publications Fund, his two books are published by the Higher Education Press and Springer Press Monograph. At the same time, Professor Cai has served as a reviewer for several important international academic journals and international academic conferences such as TKDE, ACM TOIT, IEEE Intelligent Systems, Information Science, KBS, IJCAI, AAAI, COLING, CIKM, and DASFFA.

Qing Li (Member, IEEE) received the B.Eng. degree from Hunan University, Changsha, and the M.Sc. and Ph.D. degrees from the University of Southern California, Los Angeles, all in computer science. He is currently a Chair Professor with the Department of Computing, The Hong Kong Polytechnic University. His research interests include multi-modal data management, conceptual data modeling, social media, Web services, and e-learning systems. He has authored or coauthored more than 400 publications in these areas. He is actively involved in the research community. He is a Fellow of IEE/IET, U.K., and a Distinguished Member of CCF, China. He served as a conference and program Chair/Co-Chair for numerous major international conferences. He also sits in the Steering Committees of DASFAA, ER, ACM RecSys, IEEE U-MEDIA, and ICWL. He has served as an associate editor for a number of major technical journals, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), ACM *Transactions on Internet Technology* (TOIT), *Data Science and Engineering* (DSE), *World Wide Web* (WWW), and the *Journal of Web Engineering*.