Mention Recommendation for Multimodal Microblog with Cross-attention Memory Network

Renfeng Ma[1], Qi Zhang[1], Jiawen Wang[1], Lizhen Cui[2] and Xuanjing Huang[1]
[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai, P.R.China 201203
[2]Shandong University, Jinan, Shandong Province, China
{rfma17,qz,wangjiawen16,xjhuang}@fudan.edu.cn
clz@sdu.edu.cn

ABSTRACT

The users of Twitter-like social media normally use the "@" sign to select a suitable person to mention. It is a significant role in promoting the user experience and information propagation. To help users easily find the usernames they want to mention, the mention recommendation task has received considerable attention in recent years. Previous methods only incorporated textual information when performing this task. However, many users not only post texts on social media but also the corresponding images. These images can provide additional information that is not included in the text, which could be helpful in improving the accuracy of a mention recommendation. To make full use of textual and visual information, we propose a novel cross-attention memory network to perform the mention recommendation task for multimodal tweets. We incorporate the interests of users with external memory and use the cross-attention mechanism to extract both textual and visual information. Experimental results on a dataset collected from Twitter demonstrated that the proposed method can achieve better performance than state-of-the-art methods that use textual information only.

CCS CONCEPTS

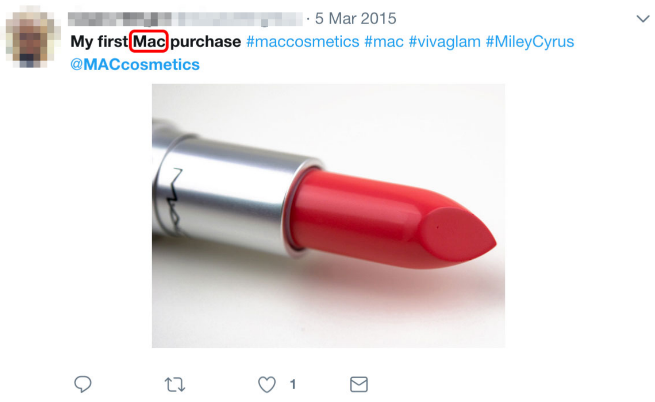
• **Information systems → Social recommendation**; **Multimedia and multimodal retrieval**;

KEYWORDS

Mention Recommendation; Multimodal Attention; Social Medias

1 INTRODUCTION

Twitter-like social media are some of the most popular and influential platforms for information generation and diffusion. According to the definition by Twitter, a tweet that contains *@username* is called a mention. In addition, if a tweet includes multiple *@username*, all of those people will see it in their own notification tabs. Hence, Twitter-like microblogging users would like to mention their friends or celebrities to report new events, promote products, share experiences, or participate in discussions. When an appropriate mention is recommended, a user could increase their exposure, promote their reputation, attract more followers, and accelerate the dissemination of information across the platform. According to the quarterly report released by Twitter[1], it had 330 million active users monthly, and the average number of followers per user was 482. Hence, it would be beneficial to have a small number of candidates when users want to mention others in a specific tweet.



[redacted] · 5 Mar 2015
My first Mac purchase #maccosmetics #mac #vivaglam #MileyCyrus @MACcosmetics

Figure 1: Example of multimodal tweet. Without visual information, we may mistakenly think MAC is an Apple laptop, whereas it is actually a lipstick by the makeup brand MAC. Hence, we should mention *@MACcosemetics* rather than *@Apple*.

Previous works have studied various aspects of the mention recommendation problem. Various supervised methods with manually constructed features like tag similarity and text similarity have been proposed to perform this task and promote tweet diffusion [21, 33]. Linguistic topic models [20, 25] and support vector machine models [31] have also been used to perform this task. Instead of trying to expand the diffusion of tweets, some works have focused on recommending a similar interest person [35–37, 40]. Because the post history of users plays quite an important role in the mention recommendation task, different kinds of resources have also been taken into consideration [5, 11, 12]. Chen et al. [5] incorporated the user's own tweet history, their retweet history, and the social relations between users to capture personal interests. Gong et al. [11]

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210026

[1] https://investor.twitterinc.com/results.cfm

treated the recommendation task as a topical translation problem with the addition of tweet content and user histories. In addition to feature engineering for machine learning models, Huang et al. [12] proposed a neural network-based method combined with the external memory of users' history. Moreover, neural network-based methods have achieved better performances than other kinds of methods.

Although some research has been done on the mention recommendation task, most of the previous methods only focused on the use of textual information. However, according to the statistics, more than 42% of tweets include more than one image[2]. Moreover, tweets with images are 150% more likely to get retweets than text-only tweets[3]. Hence, processing these multimodal tweets has become an important task. Figure 1 gives a multimodal tweet example. After reading the tweet content *"My first Mac purchase,"* we probably think the user bought a computer or laptop by Apple, which usually called a "*Mac.*" However, in this tweet, "*Mac*" should be a lipstick of the makeup brand "*Mac.*" With only textual information, it may be difficult to determine what "*Mac*" is.

To address this issue, we present a novel multimodal model to combine textual and visual information. Some previous works simply combine the text feature vector and image feature [1]. However, the correct entities or other meaningful content are often only related to a small part of the image or text. Under these conditions, using a vector to represent the image or text may lead to an incorrect final prediction as a result of the noise made by the irrelevant or unimportant part of the image or text. Motivated by work on the visual question answering task [38] and the generation of image descriptions [13], we incorporated an attention mechanism to process the textual information and visual information of a multimodal tweet. With the help of the proposed attention mechanism, our model can focus on important parts of the visual and textual information of the tweet, which can represent almost the complete meaning of the multimodal tweet. More specifically, the proposed network architecture is a neural memory network combined with a cross attention mechanism. This model can take the content of a tweet, history of its author, and history interests of candidate users into consideration, simultaneously. Meanwhile, the model can make good use of visual information by treating the image content as an assist information. Finally, predictions are calculated based on the similarity features extracted from the multimodal information of tweets, the users' histories and the candidate users' interests.

To demonstrate the effectiveness of our model, we performed experiments on a large data set collected from Twitter. The experimental results showed that the proposed method could achieve better performance than state-of-the-art methods using textual information only. The main contributions of our work can be summarized as follows.

- The mention recommendation task for multimodal tweets is novel and has not been carefully studied in previous methods. In this paper, we defined the problem and evaluated several methods for this task.
- We propose a novel cross-attention memory network that incorporates tweet-guided visual attention. It takes the content of a tweet, interests of the user, and interests of the author into consideration.
- Experimental results using a dataset constructed by us from Twitter demonstrated that our model could achieve significantly better performance than current state-of-the-art methods.

2 RELATED WORK

2.1 Mention Recommendation

Due to increasing requirements, a variety of recommendation tasks have been proposed for different problems on social media, such as content recommendation [5, 17], community recommendation [22, 39], tag recommendation [7, 10, 18], music recommendation [27], and mention recommendation [11, 12, 33]. The mention recommendation task has been studied from various aspects. Some have treated the mention recommendation task as an action to increase user's exposure and accelerate the dissemination of information across the platform. Based on this idea, Wang et al. [33] treated the task as a ranking task to find suitable users who can enhance a tweet's diffusion and incorporated several manually constructed features related to a user interest match. Zhou et al. [41] proposed a personalized ranking model that considers multi-dimensional relationships between users and mention tweets, and took the in-depth differences between mention and retweet behaviors into consideration. Li et al. [20] proposed a framework based on a linguistic topic that aims to recommend influential users and topic-cohesive interactive communities that are most relevant to the given user. In contrast, to make the tweet spread faster, some works had focused on finding the right person to be mentioned in a tweet. Li et al. [21] proposed a factor graph method to solve the mention recommendation task. A support vector machine based framework was proposed in [31], which incorporated four categories of features to solve the task. The task was also treated as a translation problem. Gong et al. [11] proposed a topical translation model incorporating the content of tweets and users' post histories to deal with this problem. Recently, neural networks have also been incorporated to perform this task. Huang et al. [12] adopted a neural network with a hierarchical attention mechanism which is the existing state-of-the-art method.

The multimedia recommendation is also related to this work and has been studied from various aspects. Chen et al. [3] addressed the problem of providing personalized video suggestions for users not only exploring the user-video graph formulated using the click-through information, but they also investigated two other useful graphs: the user-query graph and the query-video graph. Some content-based filtering models were also proposed [15]. Moreover, the group hybrid method combining collaborative and content-based recommendation models [14] was also incorporated, which focused on the performance of very Top-N recommendations. Recently, Chen et al. proposed a textual content-based attentive collaborative filtering model [4], which learns to select informative components of multimedia items, and the item-level attention module, which learns to score the item preferences.

[2] https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/

[3] https://sproutsocial.com/insights/social-media-statistics/

In view of the above descriptions, we find that most of the previous works focused only on textual information or visual information. However, both textual data and visual data contain lots of important information for the task. On the other hand, we can also clearly find that an increasing number of users prefer multi-modal information on social media platforms, as the information becomes more diverse. Therefore, we propose a cross-attention memory network that incorporates both textual and visual information to perform the mention recommendation task.

2.2 Attention Mechanisms

Attention mechanisms allow models to focus on necessary parts of inputs at each step of a task. Moreover, attention mechanism has been proved to be significantly effective in both visual related tasks and natural language processing tasks, such as machine translation [2], question answering [28, 34], image classification [23], etc. Its effectiveness results from the assumption that human recognition does not tend to process whole texts or images in their entirety. In reality, humans usually put attention into selective parts of the whole perception regions according to demand. Therefore, one important contribution of the attention mechanism is the idea of extracting important information of the inputs space, which can help the model to focus on processing the important information rather than noise and achieve a better performance on tasks.

In this work, the key idea of our model is based on the attention mechanism combined with memory network[30]. First, we use a co-attention mechanism to get textual-based new visual vectors and visual-based new tweet vectors, and using the cross-attention memory network combined with new tweet vectors and new visual vectors to fetch users' interest by extracting the important information from users' multimodal post histories. Since these important parts have been extracted, our model can achieve perfect mention recommendation.

2.3 Multimodal Tasks

As the information has become more diverse, numerous multimodal models have been proposed. Early works have usually focused on simply combining the global vectors of visual information and textual information. Recently, the task of incorporating image and text has been studied in many aspects, such as automatic image captioning task, generating descriptions for image task, and visual question answering task. Vinyals et al. [32] first extracted high-level image features and then fed them into an LSTM to generate captions. Li et al. [19] proposed a scheme to detect the copy-move forgery in an image that first segments the test image into semantically independent patches. Karpathy et al. [13] made a combination of Convolutional Neural Networks (CNNs) over image regions and bidirectional Recurrent Neural Networks (BRNNs) over sentences to generate natural language descriptions of images and their regions. Specifically, visual question answering is a major related work in these multimodal tasks. Most early works simply transfer an image captioning framework [8, 32] to visual question answering tasks [9, 26] based on CNNs and RNNs. Recently, a lot of attention mechanisms [6, 24, 42] have been proposed to align text and image information.

Different from image question answering like task that is mainly focused on extracting image features. In this work, we use the features of a given multimodal tweet captured by a co-attention mechanism to make a cross hierarchical attention of visual histories and textual histories. Hence, we model the interest relevance between the author and candidate user.

3 APPROACH

Given a tweet t_x with corresponding image t_i, its author a, and a list of candidate users U, our task is to make a decision of whether a user $u \in U$ should be recommended for the author's mention action in the tweet t. In this way, we can treat the mention recommendation task as a matching problem. Here, we preserve each author's own mention history for users as the corresponding candidate user set for the author. And a novel cross attention memory network (CAMN) architecture is proposed to perform this task. The overall architecture of the model is illustrated in Figure 2.

Firstly, we use the pre-trained VGG-Net 16 to extract the representation of images in our model, including the images of the given tweet and post histories. We then use a tweet encoder to represent the tweet. Then, using a co-attention mechanism on the representations of t_i and t_x, extracting the significant parts of the textual information and visual information of the tweet. Second, we encode the history interests of the author and the history interests of candidate user. In this part, we utilize the new representation of t_i and t_x to capture high-quality post history interest information with the help of a novel cross attention mechanism. Next, our model can formulate a high-level abstract significant representation of the given tweet t, the author a, and the candidate user u. Further, as the human cognitive learning process repeats a cognitive procedure many times, we repeat the cross attention and memory updating procedure for k steps to formulate the final representation. The steps are denoted as $k = \{0, 1, 2, \cdots, K\}$. Finally, the matched answer is predicted by a fully connected softmax layer. We describe our models in three parts. The tweet feature representation is described in Section 3.1. The cross attention memory network and matching prediction are described in Section 3.2 and Section 3.3, respectively. To be clear, we list the explanation of the key notations in Table 1.

Table 1: Annotation of symbols

N	The memory capacity
T	The maximum tweet length
M	The number of image regions
k	The layer index of cross-attention memory network

3.1 Tweet Feature Representation

Image feature representation

The image features are extracted from a pretrained 16-layer VGGNet [29]. We first rescale the images to 224 × 224 pixels and feed them into the CNNs. Rather than using a global vector as the image feature representation, we take the last pooling layer of the VGGNet to obtain the image spatial feature representation from different regions. Hence, we divide the image into 7 × 7 regions,

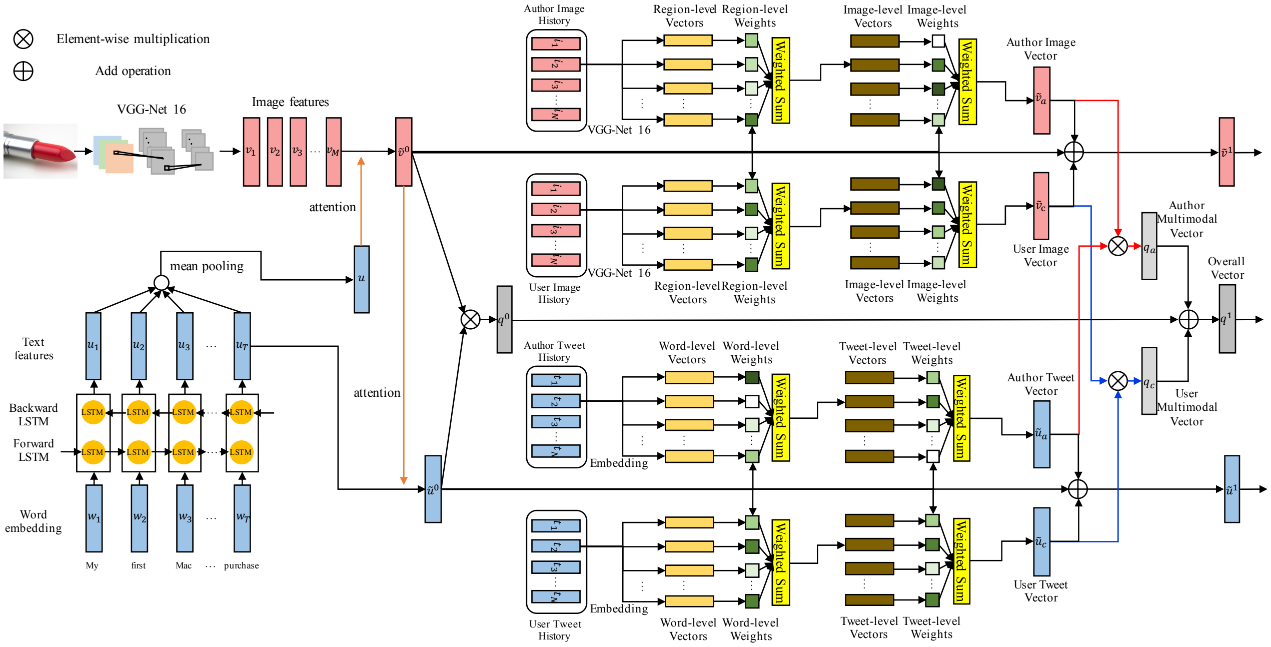


Figure 2: Overall architecture of one-layer CoA-CAMN. Our framework consisits of three components: (1) Query Tweet Modelling, (2) Tweet History Insterests Modelling and (3) Image History Insterests Modelling. Here, we denote $\widetilde{u}^0$ as the representation of query tweet and $\widetilde{v}^0$ as the representation of corresponding image, and use $\widetilde{u}^0, \widetilde{v}^0$ query cross tweet histories and image histories, respectively. q^k is the final representation of the overall architecture which modelling the interests similarity among query tweet, author and a candidate user.

and the dimension size of the feature vector for each region is 512. Therefore, an image could be represented as $v_I{}^* = \{v_i{}^* | v_i{}^* \in \mathbb{R}^D, i = 1, 2, 3, \cdots, M\}$, where $M = 7 \times 7$ is the number of image regions, and $v_i{}^*$ is a 512-dimensional feature vector for region i.

In order to make calculations more convenient, we align the dimension of each image vector to the same dimension as the tweet feature vector using a single full connection layer after the feature of the last pooling layer of VGGNet: $v_I = tanh(W_I v_I{}^* + b_I)$, where $v_I{}^*$ is the feature of last pooling layer of 16-layer VGGNet, and v_I is the image feature representation matrix after transformation by a fully connected layer.

Text feature representation

First, we transform every word w_i in a given tweet t to a one-hot vector in the size of the vocabulary. Next, we use a simple embedding layer to encode each one-hot vector to a word vector x_i distributed in a continuous space: $x_i = Mw_i$. The size of the embedding layer is $d \times |V|$, where d is the embedding dimension and $|V|$ is the size of the vocabulary. Hence, we get a word-level tweet feature representation: $t = \{x_1, x_2, \cdots, x_T\}$, where T is the maximum tweet length. More specifically, each sentence with length less than T is padded with zero vectors.

The bidirectional LSTM is a kind of RNN designed to solve the issue of ignoring future contextual information of normal RNNs. Therefore, the bidirectional LSTM is fed with each training sequence forward and backward, respectively. Hence, we utilize the bidirectional LSTM to construct a sentence-level tweet features representation. At each time step, the bidirectional LSTM unit takes the word embedding vector x_t as an input vector and outputs a hidden state h_t. The details are illustrated as follows:

$$h_t{}^{(f)} = LSTM^{(f)}(x_t, h_{t-1}{}^{(f)}), \tag{1}$$

$$h_t{}^{(b)} = LSTM^{(b)}(x_t, h_{t+1}{}^{(b)}), \tag{2}$$

where $h_t{}^{(f)}$ and $h_t{}^{(b)}$ represent the hidden states at time step t from the forward and backward LSTMs, respectively. Finally, we construct a set of text feature vectors $u_T = \{u_1, u_2, \cdots, u_T\}$ by adding the two hidden state vectors at each time step:

$$u_t = h_t{}^{(f)} + h_t{}^{(b)}, \tag{3}$$

where u_t is the representation vector of the t-th word in the context of the entire sentence. Specifically, the word embedding matrix and the bidirectional LSTMs are trained end-to-end over the whole model.

Co-attention mechanism for tweet modelling

After the above process, we get the image feature representation matrix v_I and the text feature representation matrix u_T. In view of the fact that the texts and images contain different levels of abstractions for a tweet, and the correct entities or other meaningful content are often only related to a small part of the image or text, we utilize a co-attention mechanism to generate a high-level representation of the text part and image part of a given tweet. After experiencing many kinds of attention mechanisms between tweets and images (the experiment results are recorded in Section 4), we find that the co-attention mechanism achieves perfect performance, which utilizes text-based attention and image-based attention sequentially .

Text-based visual attention

Usually, the meaning of a tweet is only related to a specific region of the corresponding image. For instance, in Figure 1, only a few parts of the image represent the "MAC" lipstick and the residual parts are white pixels. In other words, few regions of the image can be related to the tweet. Hence, instead of using a global vector to represent the image, we divide the image into 49 grids and construct an image feature matrix v_I by extracting a feature vector of each region. Then, we use a text-based attention mechanism to filter out noise and find regions that are relevant to the meaning of the corresponding text parts.

First, we use a single mean-pooling layer to summarize the sentence-level representation of a given tweet: $\overline{u} = \frac{1}{T}\sum_{i=1}^{T} u_i$. Next, we incorporate image attention with the help of the sentence-level representation:

$$h_M = tanh(W_{v_M} v_M) \odot tanh(W_{v_U} \overline{U}), \tag{4}$$

$$a_M = softmax(W_h h_M), \tag{5}$$

where $\overline{U} \in \mathbb{R}^{d \times M}$ is a matrix formulated by M columns of $\overline{u}$ and $v_M \in \mathbb{R}^{d \times M}$, d is the dimension of the representation, and M is the number of divided regions of each image. We use $\odot$ to denote the element-wise multiplication of two matrices.

Since the attention probability a_m of each image region m is calculated from the above equation, the new representation of the image is formulated as the weighted sum of the image region vectors.

$$\widetilde{v}_I = \sum_m a_m v_m \tag{6}$$

Image-based textual attention

However, the text-based visual attention can make the model focus on those important image regions. Following the same idea, we use image-based textual attention to help model which word is more important, which formulates the sentence-level meaning of a tweet. The process of image-based textual attention is similar to text-based visual attention. To be specific, we utilized the new image representation vector $\widetilde{v}_I$ to query the original textual feature u_T, generating the a new text representation $\widetilde{u}_T$ based on the textual attention probability distributions. The detail is illustrated as follows:

$$z_T = tanh(W_{u_T} u_T) \odot tanh(W_{u_V} \widetilde{V}_I), \tag{7}$$

$$a_T = softmax(W_z z_T), \tag{8}$$

where $\widetilde{V}_I \in \mathbb{R}^{d \times T}$ is a matrix formulated by T columns of $\widetilde{v}_I$, $u_T \in \mathbb{R}^{d \times T}$and T is the max length of tweets. And "$\odot$" denotes the element-wise multiplication of two matrices.

As the attention probability a_t of the t-th word is calculated, the new representation of the tweet is formulated by the weighted sum of each word vector:

$$\widetilde{u}_T = \sum_t a_t u_t \tag{9}$$

3.2 Cross Attention Memory Network

Obviously, a user's post histories can be used to model the user's history interests. And, both images and tweets contain important information that can be extremely helpful to generate a perfect mention prediction. Hence, we propose a cross attention memory network to model the similarity of multimodal history interests between the author and the candidate user. It is obvious that the tweet histories and the corresponding image histories stored in the memory have a hierarchical structure. For instance, each image document has many images: $D_I = \{e_1, e_2, \cdots, e_N\}$ and an image-level structure. Each image has been divided into regions: $e = \{v_1, v_2, \cdots, v_M\}$ and a region-level structure. Each tweet document has many tweets: $D_T = \{t_1, t_2, \cdots, t_N\}$ and a tweet-level structure. Each tweet also has many words: $t = \{w_1, w_2, \cdots, w_T\}$. Moreover, not all tweets and images in the history memory contain equally relevant information for modelling the history interests, and not all regions in an image or all words in a tweet are equally meaningful. Hence, we propose a hierarchical architecture to model a user's image history interests and tweet history interests, respectively.

When given a query tweet and a corresponding image, we constructed a final representation after k-layer cross attention memory network. We denote the above new tweet feature representation $\widetilde{u}_T$ as $\widetilde{u}^0$ and the corresponding new image feature representation $\widetilde{v}_I$ as $\widetilde{v}^0$. Then, we utilize $\widetilde{u}^k$ to formulate textual attention probabilities cross the author's tweet histories and candidate user's tweet histories, and construct the $\widetilde{u}^{k+1}$. Then $\widetilde{v}^k$ is used to construct a visual attention cross-reference between the author's image histories and candidate user's image histories, and get updated the same as $\widetilde{u}^k$. And the final vector q_k is used to predict mention action.

Region-level encoder

Given an input image set $D_I = \{e_1, e_2, \cdots, e_N\}$, each region's original representation $r_{i,j} \in e_i$ is formulated by a 16-layer VGGNet and saved as a visual memory vector. The memory vector $v_{i,j}$ in this step is called input memory, which projects the input history image regions into the same space. The dimension of $r_{i,j}$ is 512. In order to make calculation more convenient, we align the dimension of each region's original vector $r_{i,j}$ to the same dimension as the tweet feature vector by using a single full connection layer on the region's original vectors $v_{i,j} = Wr_{i,j}$.

With an underlying intuition that not all regions in each image are equally relevant for modelling the image history interests, at k-th layer of cross-attention network, we utilize the last step of the image representation vector $\widetilde{v}^{k-1}$ to query a user's region vector set, generating the representation of each history image based on the region attention probability distributions. The detail is illustrated as follows:

$$h^k_{i,M} = tanh(W^k_M v_{i,M}) \odot tanh(W^k_{\widetilde{v}} \widetilde{V}^{k-1}_M), \tag{10}$$

$$a^k_{i,M} = softmax(W^k_h h^k_{i,M}), \tag{11}$$

$$v^*_i = \sum_j^M a^k_{i,j} v_{i,j}, \tag{12}$$

where $\widetilde{V}^{k-1}_M \in \mathbb{R}^{d \times M}$ is a matrix formulated by M columns of $\widetilde{v}^{k-1}$ and M is the region number of each image.

Image-level encoder

As in the above descriptions, we get a new representation v^*_i for each history image based on a region-level attention mechanism. However, it is obvious that not every image is equally relevant to constructing a user's image history interests. Hence, in order to model the whole image history interests of a user, we utilize the last step of the image representation vector $\widetilde{v}^{k-1}$ to query the new representations of each history image. Modelling the representation of a user's image histories based on the image-level attention probability distributions:

$$h_N^k = tanh(W_N^k v_N^*) \odot tanh(W_{\widetilde{v}_N}^k \widetilde{V}_N^{k-1}), \tag{13}$$

$$a_N^k = softmax(W_{h_N}^k h_N^k), \tag{14}$$

$$\widetilde{v}_* = \sum_i^N a_i^k v_i^*, \tag{15}$$

where $\widetilde{V}_N^{k-1} \in \mathbb{R}^{d \times N}$ is a matrix formulated by N columns of $\widetilde{v}^{k-1}$, and N is the image amount of each image history document.

Therefore, through the above steps, we get the representation $\widetilde{v}_* \in \mathbb{R}^d$ for a user's whole image history interests ("$*$" can represent an author or a candidate user) and d is the dimension of the representation. In other words, we model the image history interests of the author as $\widetilde{v}_a$ and the image history interests of the candidate user as $\widetilde{v}_c$ by following the above steps, respectively.

Word-level encoder

Similar to the hierarchical architecture of the image history, the text history also has a two-level architecture. Give an tweet history set $t_1, t_2, ..., t_N$, first, each word $w_{i,j}$ of t_i is embedded into a textual memory vector $c_{i,j}$ (dimension of $c_{i,j}$ is d) using an embedding matrix A (size of A is $d \times |V|$), as $c_{i,j} = Aw_{i,j}$. The memory vector $c_{i,j}$ in this process is similar to the image memory vector $v_{x,y}$, which is called input memory and projects the input words of historical tweets into the same space.

Leveraging the advantage of filtering irrelevant words, at the k-th cross-attention layer, we use the last step of the tweet representation vector $\widetilde{u}^{k-1}$ to generate attention probabilities over a user's word memory vector set.The match between input memory vector $c_{i,j}$ and $\widetilde{u}^{k-1}$ is computed by incorporating the inner product followed by a softmax layer:

$$z_{i,T}^k = (\widetilde{u}^{k-1})^{tr} c_{i,T}, \tag{16}$$

$$p_{i,T}^k = softmax(W_z^k z_{i,T}^k), \tag{17}$$

where $(\widetilde{u}^{k-1})^{tr}$ is the transpose of last step of the tweet representation vector $\widetilde{u}^{k-1}$ and T is the maximum length of each tweet.

Different from the region-level encoder, we use a new embedding matrix B to embed each word $w_{i,j}$ into another word memory vector $u_{i,j}$ (of dimension d and named output memory), as $u_{i,j} = Bw_{i,j}$. Finally, the representation of the tweet t_i is obtained by summing all output memory vectors weighted by the above attention probability:

$$u_i^* = \sum_j^T p_{i,j}^k u_{i,j}, \tag{18}$$

Following the above process, each tweet in the tweet history document is converted into a fixed-length vector that represents the interest embedding of the tweet.

Tweet-level encoder

In order to model the complete tweet history interests of a user, we propose a tweet-level encoder to aggregate important parts of the tweet history document. Given the encoded set of tweets $s = \{u_1^*, u_2^*, \cdots, u_N^*\}$, the history interest representation of the tweet history document is formed by a weighted sum of these tweet representations. The weights over the each tweet are interpreted as the importance level of a particular tweet in the document. The equation of this procedure is as follows:

$$z_N^k = tanh(W_N^k u_N^*) \odot tanh(W_{\widetilde{u}_N}^k \widetilde{U}_N^{k-1}), \tag{19}$$

$$p_N^k = softmax(W_{z_N}^k z_N^k), \tag{20}$$

$$\widetilde{u}_* = \sum_i^N p_i^k u_i^*, \tag{21}$$

where $\widetilde{U}_N^{k-1} \in \mathbb{R}^{d \times N}$ is a matrix formulated by N columns of $\widetilde{u}^{k-1}$ and N is the tweet amount of a tweet history document.

Just as the image history interests representation, the label "$*$" can represent the author or the candidate user in the tweet history interests representation $\widetilde{v}_* \in \mathbb{R}^d$. In this way, we model the tweet history interests of the author as $\widetilde{u}_a$ and the tweet history interests of a candidate user as $\widetilde{u}_c$ by following the above steps, respectively.

Stacked cross-attention network

For modelling more complex history interests, the similarity between the author and a candidate user, according to a query tweet, we can try to repeat the cross-attention memory network iteratively by using the newly generated representations. Formally, the stacked procedure can be summarized as follows: for the k-th (where k is greater than or equal to 1) cross-attention layer, we construct the image history interests and the tweet history interests representation for the author and a candidate user based on the query tweet, respectively. The new query vector is formed by adding the new feature vector to the previous vector, the detail is as follows:

$$\widetilde{u}^k = \widetilde{u}^{k-1} + \widetilde{u}_a^k + \widetilde{u}_c^k, \tag{22}$$

$$\widetilde{v}^k = \widetilde{v}^{k-1} + \widetilde{v}_a^k + \widetilde{v}_c^k, \tag{23}$$

where $\widetilde{u}^0$ is initialized by tweet presentation $\widetilde{u}_T$ of the query tweet, and $\widetilde{v}^0$ is initialized by image presentation $\widetilde{v}_I$ of the query tweet.

Further, the k-th global representation vector, which is used to model the history interests similarity between author and a candidate user according to a query tweet, is updated after the k-th image history interests and tweet history interests representation for the author and a candidate user, respectively:

$$q_a^k = \widetilde{u}_a^k \odot tanh(W_a \widetilde{v}_a^k), \tag{24}$$

$$q_c^k = \widetilde{u}_c^k \odot tanh(W_c \widetilde{v}_c^k), \tag{25}$$

$$q^k = q^{k-1} + q_a^k + q_c^k, \tag{26}$$

where q_a^k is the whole interests representation (named final presentation) of the author's tweet histories and the corresponding image histories at k-th cross-attention memory layer, and q_c^k is the representation of the candidate user. Particularly, considering the inequalities of information density of word vectors and picture region vectors, we apply an additional layer W_a, W_c to make the image history representation and tweet history representation have similar information density.

Table 2: Comparison results on the testing dataset. We divided the compared approaches into three categories based on different mechanisms. Category I belongs to traditional machine learning methods. Category II is based on deep neural networks. Category III includes different variants of our approach.

Method		Precision	Recall	F-Score	MRR	Hits@3	Hits@5
I	NB	0.503	0.477	0.490	0.614	0.653	0.744
	PMPR [21]	0.650	0.624	0.637	0.742	0.793	0.858
	CAR [31]	0.685	0.665	0.675	0.770	0.817	0.876
II	LSTM+CNN	0.550	0.532	0.541	0.626	0.676	0.712
	MLAN [38]	0.735	0.677	0.705	0.807	0.821	0.846
	DAN [24]	0.772	0.748	0.760	0.809	0.834	0.850
	AU-HMNN [12]	0.771	0.751	0.761	0.828	0.852	0.902
III	ITA-CAMN	0.754	0.732	0.743	0.817	0.848	0.899
	TVA-CAMN	0.811	0.786	0.798	0.858	0.879	0.916
	CoA-CAMN	**0.817**	**0.793**	**0.804**	**0.862**	**0.880**	**0.919**

3.3 Prediction

Finally, based on the final presentation obtained from the above process, we utilize a single-layer softmax classifier to determine whether or not a candidate user should be mentioned to the author according to a query tweet:

$$f = \sigma(W_q q^k + b_q), \tag{27}$$

where W_q, b_q are parameters of a hidden layer, q^k is the final representation obtained after the k times cross-attention layer and σ is a non-linear activation function.

The final prediction is made by the following equations:

$$p(y = i|f; \theta_s) = \frac{exp(\theta_s^i f)}{\sum_j exp(\theta_s^j f)}, \tag{28}$$

where θ_s^i is a weight vector of the i-th class and $j \in \{0, 1\}$.

In our work, the training objective function is formulated as follows:

$$J = \sum_{(t_q, a, c, i) \in D} -logp(i|t_q, a, c; \theta), \tag{29}$$

where D is the training set. $i \in \{0, 1\}$ is the label of the triple (t_q, a, c), and when $i = 1$, the candidate user c should be recommended to the author a's "@" action in the tweet t_q, and $i = 0$ represents the candidate user c that should not be recommended. θ is the whole parameter set of our model.

To minimize the objective function, we use a stochastic gradient descent (SGD) with the Adam [16] update rule. Moreover, we use the dropout and add l_2-norm terms for the regularization (the parameter of l_2-norm terms is 9×10^{-9} in our training procedure).

4 EXPERIMENT

4.1 Baseline

To analyze the effectiveness of our model, we evaluated some traditional and state-of-the-art methods as baselines as follows on the constructed corpus:

- **NB**: Naive Bayes is implemented with bag-of-word features transformed from the posting history.
- **PMPR**: The Personalized Mention Probabilistic Ranking (PMPR) system is proposed in [21] to solve the mention recommendation problem.
- **CAR**: The Context-aware At Recommendation (CAR) model is a ranking support vector machine model proposed in [31] to locate the target users.
- **LSTM+CNN**: LSTM+CNN is normally combined the textual feature processed by LSTM with visual feature processed by CNN to make a prediction.
- **MLAN**: Multi-level Attention Networks (MLAN) is proposed in [38] to deal with the visual question answering task. It is incorporated to evaluate the effectiveness of our model to deal with the multimodal task.
- **DAN**: Dual Attention Networks (DAN) is proposed in [24] and it perform well in visual question answering task and image-text matching task. The reason for incorporating this model is the same as for MLAN.
- **AU-HMNN**: AU-HMNN is proposed in [12], which incorporates only textual information of query tweets and users' history. This was the state-of-the-art approach used for the mention recommendation task.

Table 3: Statistics of the evalution dataset.

#Tweets	200,465
#Images	200,465
#Users	15,539
#Avg.Mention / Author	39.45
#Avg.Mentioned User/Author	9.51

4.2 Dataset and Setup

To evaluate the effectiveness of our proposed model, we collected public tweets to construct a dataset from Twitter. Firstly, we randomly selected 4,000 users as the authors and crawled their post histories. And the collection contained 15.3 million tweets. Then, from these tweets, we extracted those contained both images and at least one @username, and collected the corresponding mentioned

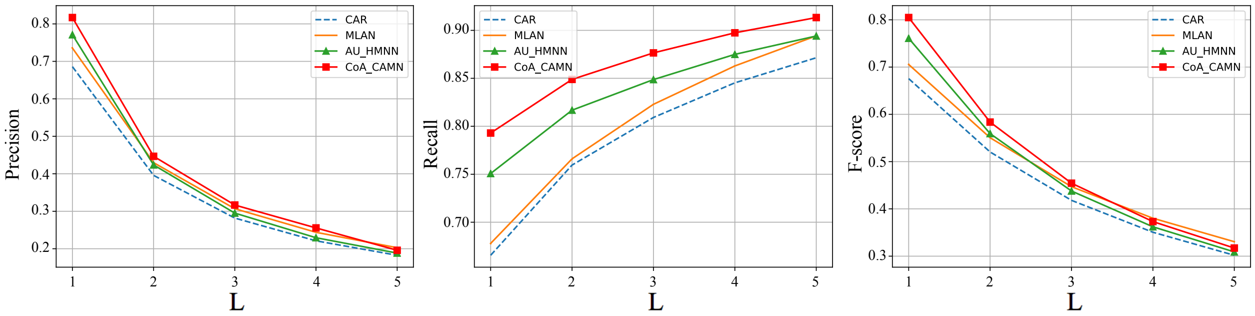


Figure 3: Precision, Recall and F-score with different amount of recommended users

users. In this step, 3,112 authors and a total of 122,770 query tweets were extracted. The amount of mentioned users was 12,427. Next, we crawled the mentioned users' histories, and 131.6 million tweets were collected. Finally, we also selected those tweets that contained images from the mentioned users' histories. The detailed statistics are shown in Table 3, the average number of mention behaviours per central author was 39.45, and the average number of users that the central authors mentioned was 9.51. For each query tweet, the mentioned history of each author was considered as a candidate. We split the dataset into a training and a testing set with an 80/20 ratio, and randomly selected 20% of the training set as the development set.

In this work, we filtered out the stop words and low-frequency words for texts. For images, we downloaded images from the retrieved urls and rescaled them to 224 × 224. Then we used a pre-trained 16-layer VGGNet to extract features. The outputs of the last pooling layer of the VGGNet were extracted as the image features. For the memory portion, the capacity of the memory was restricted to 5 tweets with corresponding images, and the maximum length of each tweet was 31. In other words, we randomly extracted 5 tweets from each author and each mentioned user history, and used these 5 tweets to present their history interests and stored them in the supporting memory. The embedding dimension in the experiment was 300 (we also transferred the image feature dimension from 512 to 300), and the depth of cross-attention memory layer was set to 7. The learning rate was set to 0.01, and the dropout rate was set to 0.2.

Further, we used precision, recall, and F-score to evaluate the results, and incorporated Hits@3 and Hits@5 to represent the percentage of correct results recommended from the top L results. Moreover, we also used the Mean Reciprocal Rank (MRR) metrics to evaluate the rank of the recommended results.

4.3 Results and Discussion

The performance of different methods on our dataset is listed in Table 2. The results in all the metrics were obtained when we recommended the top one candidate user for each tweet. We can observe that our proposed model (CoA-CAMN) achieves a better performance than other comparison methods of all metrics.

Compared with AU-HMNN, which was the state-of-the-art method for the mention recommendation task, the proposed model (CoA-CAMN) achieves a relative improvement of 6.0% in precision, along with a 5.6% increase in recall and 5.7% increase in F-score. Moreover, the best results of our proposed model for Hits@3 and Hits@5 are greater than 0.882 and 0.920, respectively. In other words, 88.2% of correct users can be found in the top 3 recommendation list and 92.0% of the users can be recommended in the top 5. Particularly, the MRR results of CoA-CAMN are also better than other methods, which illustrates that the rank of the result predicting a better recommendation of candidate users.

In order to prove the effectiveness of incorporating users' tweet posting histories and corresponding images, we also used MLAN and DAN on our dataset. Moreover, these methods were the state-of-the-art methods for visual question answering task and also performed well in other multimodal tasks. From the results table, we can observe that our proposed model (CoA-CAMN) consistently achieves a better performance than these multimodal models in all evaluation results. Compared with the DAN (performs better than MLAN in our dataset), our model achieves more than 5.7% relative improvements in precision, recall and F-score. Particularly, the best results of our proposed model for MRR and Hits@5 are relatively greater than 6.5% and 8.2%, respectively. Hence, by incorporating users' tweet posting histories and corresponding images, our proposed model did indeed perform well on the mention recommendation task.

Category III is a comparison of the results of Image-based Textual Attention-Cross Attention Memory Network (ITA-CAMN), Text-based Visual Attention-Cross Attention Memory Network (TVA-CAMN) and Coattention-Cross Attention Memory Network (CoA-CAMN). The comparison shows that CoA-CAMN achieves a better result than other two variants of our proposed model. As TVA-CAMN can only use text information to generate visual attention distribution to model the query tweet and the corresponding image, we find that co-attention mechanism is beneficial to modelling both the query tweet and the corresponding image. We believe the importance of the tweet and the corresponding image for summarizing global information is not equal, which limits the performance of the other variant of our proposed model (ITA-CAMN). From the results of CoA-CAMN, TVA-CAMN and ITA-CAMN, we can observe that

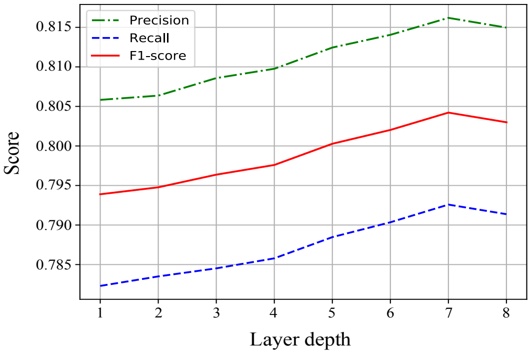


Figure 4: Performance on different layer of cross-attention memory network

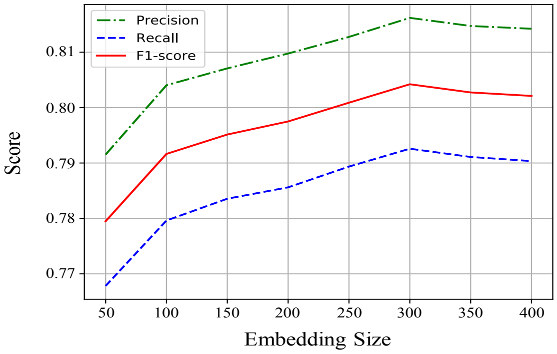


Figure 5: Influence of Embedding size

treating the tweet history as a guiding part and the image history as assisting part can significantly improve the performance.

Figure 3 also shows the Precision, Recall and F-score of the models with different numbers of recommended users. Each point of the curve represents the number of users recommended ranging from 1 to 5. In order to make the results more clear, we selected some representative methods and used the red line combined with a square label to draw the figure. Obviously, Precision decreases and Recall increases as the number of recommended users increases, except when five users are recommended, at which our model is slightly lower than MLAN in Precision. Particularly, our model achieves extremely better performance than other methods in Recall. Moreover, we obtain the highest F-score when recommending the top one user. The curve that is higher on the graph indicates the better performance. From the figure, we can see that the performance of our proposed model is the highest of all the methods when the number of users recommended ranges from 1 to 4. The most important is that the proposed method was significantly better than the state-of-the-art methods.

4.4 Parameter Influence

The proposed model contains several critical hyper-parameters. We analyzed the influence of critical parameters from the following perspectives: 1) the depth of cross-attention layer, 2) the embedding dimension, and 3) the dropout. We varied one parameter and fixed the others in turn to evaluate their influences. Based on the experimental results, we can observe that the proposed model could achieve stable performance, in the condition of various parameter settings.

The first parameter we evaluated is the depth of cross-attention layer, which we varied from 1 to 8 in this experiment. In Figure 4, we draw the Precision, Recall and F-score curves to show the depth of the cross-attention layer's influence on the performance. Along with the increase in the depth of the cross-attention layer, the results are better. We also obtained the best performance with the 7-layer cross-attention, which indicates the robustness of our model along with the depth deeper. Since increasing the number of layers of the network will make the model more complex with more parameters, and the result is not significantly improved, hence, the performance of 8-layer cross-attention is slightly lower than 7-layer. The figure demonstrates that the multi-depth cross-attention works better than single-level attention.

The second parameter is the embedding dimension. To evaluate how it influenced the performance, we fixed the depth of the cross-attention layer to 7 and tried different embedding dimensions. The comparison results shown in Figure 5 demonstrate that the models with a high embedding dimension performed better than those with a low dimension. The results improved when the dimension was increased from 50 to 300, and the result of the 400 embedding dimension was slightly lower than the 300 dimension. This shows that if we want to give more effective suggestions, the 300 embedding dimension would be a good choice.

Third, we compared the performance of our model with and without dropout layers, and the results are shown in Table 4. The depth of the cross-attention layer was 7 and the embedding dimensions was 300. It is obvious that the model achieve better performance with the help of dropout layer. Although without drop out layer, our model achieved greater than 4.1% relative improvements in each category, compared with the state-of-art method.

Table 4: Performance with and without dropout on our datasets

With Dropout	Precision	Recall	F-score	MRR
No	0.804	0.781	0.792	0.853
Yes	**0.817**	**0.793**	**0.804**	**0.862**

5 CONCLUSION

Due to the dramatically increase in social media use and the diversity of information, in this paper we proposed and studied a novel task for recommending users for multimodal microblogs to improve the usability of the user experience on mention actions. We proposed CoA-CAMN to combine textual and visual information, as well as modeled users' posting history interests by applying a novel attention mechanism on external memory. Since tweets and images are not equally important in modelling query tweet representation, we utilize the co-attention network, which generates textual attention and visual attention sequentially. We used the representation of a query tweet and corresponding images combined with a cross-attention mechanism to query the posting histories between the author and a candidate user. We also constructed a large data collection retrieved from live microblog services to evaluate the effectiveness of our model. Experimental results showed that the proposed method achieves better performance than state-of-the-art methods using textual information only.

ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 61532011, 61473092, and 61472088), STCSM (No.16JC1420401,17JC1420200), National Key R&D Program 2017YFB1400102.

REFERENCES

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. VQA: Visual Question Answering. *International Journal of Computer Vision* 123, 1 (2015), 1–28.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *Computer Science* (2014).

[3] Bisheng Chen, Jingdong Wang, Qinghua Huang, and Tao Mei. 2012. Personalized video recommendation through tripartite graph propagation. In *ACM International Conference on Multimedia*. 1133–1136.

[4] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–344.

[5] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. 2012. Collaborative personalized tweet recommendation. (2012), 661–670.

[6] Kan Chen, Jiang Wang, Liang Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering. *Computer Science* (2015).

[7] Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning Topical Translation Model for Microblog Hashtag Suggestion. In *International Joint Conference on Artificial Intelligence*. 2078–2084.

[8] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2014. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39, 4 (2014), 677.

[9] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*. 2296–2304.

[10] Yuyun Gong and Qi Zhang. 2016. Hashtag recommendation using attention-based convolutional neural network. In *International Joint Conference on Artificial Intelligence*. 2782–2788.

[11] Yeyun Gong, Qi Zhang, Xuyang Sun, and Xuanjing Huang. 2015. Who Will You "@"?. In *ACM International on Conference on Information and Knowledge Management*. 533–542.

[12] Haoran Huang, Qi Zhang, Xuanjing Huang, Haoran Huang, Qi Zhang, and Xuanjing Huang. 2017. Mention Recommendation for Twitter with End-to-end Memory Network. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*. 1872–1878.

[13] Andrej Karpathy and Fei Fei Li. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39, 4 (2017), 664.

[14] Ondrej Kaššák, Michal Kompan, and Mária Bieliková. 2016. Personalized hybrid recommendation for group of users: top-N multimedia recommender. *Information Processing & Management* 52, 3 (2016), 459–477.

[15] Jae Kyeong Kim, Hyea Kyeong Kim, and Yoon Ho Cho. 2008. A user-oriented contents recommendation system in peer-to-peer architecture. *Expert Systems with Applications* 34, 1 (2008), 300–312.

[16] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

[17] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M. Jose. 2009. On social networks and collaborative recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 195–202.

[18] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. 2009. Latent dirichlet allocation for tag recommendation. In *ACM Conference on Recommender Systems, Recsys 2009, New York, Ny, Usa, October*. 61–68.

[19] Jian Li, Xiaolong Li, Bin Yang, and Xingming Sun. 2017. Segmentation-Based Image Copy-Move Forgery Detection Scheme. *IEEE Transactions on Information Forensics & Security* 10, 3 (2017), 507–518.

[20] Lei Li, Wei Peng, Saurabh Kataria, Tong Sun, and Tao Li. 2013. FRec: a novel framework of recommending users and communities in social media. In *ACM International Conference on Conference on Information & Knowledge Management*. 1765–1770.

[21] Quanle Li, Dandan Song, Lejian Liao, and Li Liu. 2015. Personalized Mention Probabilistic Ranking–Recommendation on Mention Behavior of Heterogeneous Social Network. In *International Conference on Web-Age Information Management*. Springer, 41–52.

[22] Hao Ma, Irwin King, and Michael R. Lyu. 2009. Learning to recommend with social trust ensemble. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 203–210.

[23] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. 3 (2014), 2204–2212.

[24] Hyeonseob Nam, Jung Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. (2017).

[25] Marco Pennacchiotti and Siva Gurumurthy. 2011. Investigating topic models for social media user recommendation. In *International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April*. 101–102.

[26] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in neural information processing systems*. 2953–2961.

[27] Markus Schedl and Dominik Schnitzer. 2013. Hybrid retrieval approaches to geospatial music recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 793–796.

[28] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Computer Science* (2014).

[30] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-To-End Memory Networks. *Computer Science* (2015).

[31] Liyang Tang, Zhiwei Ni, Hui Xiong, and Hengshu Zhu. 2015. Locating targets through mention in Twitter. *World Wide Web-internet & Web Information Systems* 18, 4 (2015), 1019–1049.

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[33] Beidou Wang, Can Wang, Jiajun Bu, Chun Chen, Wei Vivian Zhang, Deng Cai, and Xiaofei He. 2013. Whom to mention: expand the diffusion of tweets by@ recommendation on micro-blogging systems. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 1331–1340.

[34] Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604* (2016).

[35] Ke Xu, Yi Cai, Huaqing Min, Xushen Zheng, Haoran Xie, and Tak Lam Wong. 2017. UIS-LDA: a user recommendation based on social connections and interests of users in uni-directional social networks. In *the International Conference*. 260–265.

[36] Ke Xu, Xushen Zheng, Yi Cai, Huaqing Min, Zhen Gao, Benjin Zhu, Haoran Xie, and Tak Lam Wong. 2017. Improving User Recommendation by Extracting Social Topics and Interest Topics of Users in Uni-Directional Social Networks. *Knowledge-Based Systems* (2017).

[37] Rui Yan, Mirella Lapata, and Xiaoming Li. 2012. Tweet recommendation with graph co-ranking. In *Meeting of the Association for Computational Linguistics: Long Papers*. 516–525.

[38] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017. Multi-level Attention Networks for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 4187–4195.

[39] Wei Zhang, Jianyong Wang, and Wei Feng. 2013. Combining latent factor model with location features for event-based group recommendation. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 910–918.

[40] Gang Zhao, Mong Li Lee, Wynne Hsu, Wei Chen, and Haoji Hu. 2013. Community-based user recommendation in uni-directional social networks. In *ACM International Conference on Information & Knowledge Management*. 189–198.

[41] Ge Zhou, Lu Yu, Chu Xu Zhang, Chuang Liu, Zi Ke Zhang, and Jianlin Zhang. 2016. A Novel Approach for Generating Personalized Mention List on Micro-Blogging System. In *IEEE International Conference on Data Mining Workshop*. 1368–1374.

[42] Yuke Zhu, Oliver Groth, Michael Bernstein, and Fei Fei Li. 2016. Visual7W: Grounded Question Answering in Images. In *Computer Vision and Pattern Recognition*. 4995–5004.