AMNN: Attention-Based Multimodal Neural Network Model for Hashtag Recommendation

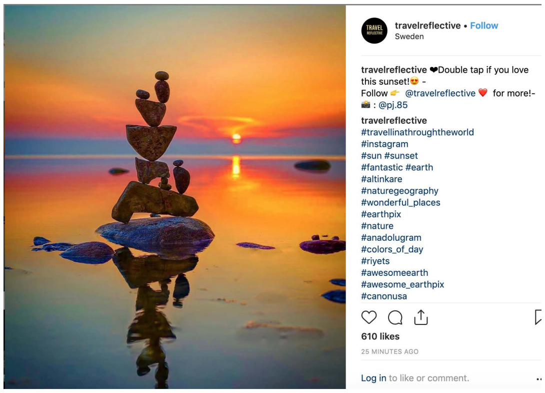
Qi Yang, Gaosheng Wu, Yuhua Li, *Member, IEEE*, Ruixuan Li, *Member, IEEE*,
Xiwu Gu, Huicai Deng, and Junzhuang Wu

***Abstract*—In the real-world social networks, hashtags are widely applied for understanding the content of an individual microblog. However, users do not always take the initiative in attaching hashtags when posting a microblog so that much effort has been invested for automatically hashtag recommendation. As a new trend, users no longer only post texts but prefer to share with multimodal data, such as images. To deal with these situations, we propose an attention-based multimodal neural network model (AMNN) to learn the representations of multimodal microblogs and recommend relevant hashtags. In this article, we convert the hashtag recommendation task into a sequence generation problem. Then, we propose a hybrid neural network approach to extract the features of both texts and images and incorporate them into the sequence-to-sequence model for hashtag recommendation. Experimental results on the data set collected on Instagram and two public data sets demonstrate that the proposed method outperforms state-of-the-art methods. Our model achieves the best performance in three different metrics: precision, recall, and accuracy. The source code of this article can be obtained from "https://github.com/w5688414/AMNN."**

***Index Terms*—Attention mechanism, hashtag recommendation, multimodal data, neural network, sequence-to-sequence.**

I. Introduction

SOCIAL networks have evolved into powerful platforms for people to communicate and share information in recent years. Along with the growth of mobile Internet, people are engaged in online activities and spend more time on social media, resulting in a massive volume of data generated. To avoid being overwhelmed, a good choice that improves information diffusion is through the use of hashtags. As a manually user-driven tagging mechanism, hashtags can be used for marking keywords or topics within per microblog. Users create and use hashtags by prefixing the hash symbol (#) in front of a word or a short phrase, e.g., #sunset. Hashtag provides a way to organize the user-generated data easily and makes the information more accessible. Many previous works have demonstrated the effectiveness of hashtags in social data mining, such as information retrieval [1], sentiment analysis [2], event detection, and topic tracking [3], because hashtags are more inclusive and informative. However, only a portion of microblogs contains hashtags. For example, only 24% of tweets contain at least one hashtag[1] on Twitter while over 500 million tweets being posted per day. Hence, the task of hashtag recommendation in social networks has drawn more and more attention in recent years.



travelreflective • Follow
Sweden

travelreflective ❤Double tap if you love this sunset!😍 -
Follow 👉 @travelreflective ❤️ for more!-
📸 : @pj.85

travelreflective
#travellinathroughtheworld
#instagram
#sun #sunset
#fantastic #earth
#altinkare
#naturegeography
#wonderful_places
#earthpix
#nature
#anadolugram
#colors_of_day
#riyets
#awesomeearth
#awesome_earthpix
#canonusa

610 likes
25 MINUTES AGO

Log in to like or comment.

Fig. 1. Example of a short microblog with images as well as the corresponding hashtag list.

Although there are a lot of works devoted to this task, most of the existing approaches only focus on the text information [4]–[10]. However, a new trend has emerged in social networks where people tend to attach pictures with texts when sharing their ideas. It can be counted that more than a third of all microblogs contain not only texts but also images [11], [12] in the microblog platforms. On the other hand, with the rising of photograph and video sharing social services, such as Instagram and Flicker, more multimodal microblogs with hashtags are available. Hence, multimodal hashtag recommendation becomes a new challenging task. An example of a microblog with images and texts as well as the corresponding hashtag list can be seen in Fig. 1. It is worth noticing that hashtags, such as #earth and #nature, are also relevant to the picture and cannot be extracted from the texts directly. Hence, better hashtags could be recommended by jointly considering the text and image information for multimodal microblogs.

Manuscript received June 26, 2019; revised October 8, 2019 and January 10, 2020; accepted March 24, 2020. Date of publication April 30, 2020; date of current version June 10, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2016QY01W0202 and Grant 2016YFB0800402 and in part by the National Natural Science Foundation of China under Grant U1936108, Grant U1836204, Grant 61572221, Grant 61572222, and Grant 61502185. *(Corresponding authors: Yuhua Li; Ruixuan Li.)*

The authors are with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: ayang7@hust.edu.cn; wugaosheng@hust.edu.cn; idcliyuhua@hust.edu.cn; rxli@hust.edu.cn; guxiwu@hust.edu.cn; hcd@hust.edu.cn; wujunzhuang@hust.edu.cn).

Digital Object Identifier 10.1109/TCSS.2020.2986778

[1] https://www.omnicoreagency.com/twitter-statistics/

Along with the impressive achievements in computer version and neural language processing (NLP) tasks, various approaches based on the deep neural network have been proposed for hashtag recommendation [13]–[16]. These models, which adopt the neural network and attention mechanism to extract features from microblogs, have achieved promising results recently. However, most of these studies still perform the hashtag recommendation tasks based on textual information only. Automatically generating a list of hashtags based on images is a very challenging task because hashtags are not only related to the content of the images but also the latent semantic information. The correlations between hashtags and images can be called weak correlations. Thus, previous methods using only textual features cannot be directly used for this task. Meanwhile, deep learning methods have achieved dominant performance on neural image caption (NIC) generation problems [17]–[22], which is a straightforward task that automatically generates textual description for an image by a neural network. To this end, the coattention network [11] is presented to combine textual and visual information together to recommend hashtags for multimodal tweets. However, as the coattention mechanism generates textual and visual attentions in a coguided way, the coattention network cannot be well employed to image-only and text-rarely situations, which are common in current social media services. Moreover, the coattention network model uses a multiclass softmax classifier with cross-entropy loss to conduct the hashtag recommendation task, leading to the performance decline in multiple hashtag environments.

Different from the previous research, we investigate the correlations among hashtags and convert the hashtag recommendation task into a sequence prediction problem. The intuition behind is based on the observation that a single hashtag cannot describe all the content of one microblog comprehensively, and all hashtags together can constitute the complete information, as shown in Fig. 1. In other words, implicit correlations between two hashtags may exist. To this end, we propose a novel attention-based multimodal neural network (AMNN) model to capture the latent interactions among images, texts, and hashtags, in which the encoder–decoder architecture is employed for hashtag sequence prediction. Inspired by the achievements in the image captioning task, we adopt a hybrid neural network model to extract features from multimodal data with the attention mechanism.

Particularly, the hashtag recommendation task that we focus on in this article is reinterpreted as a sequence generation problem. We employ the softmax+seq2seq mechanism to achieve the expected effect. Besides, considering that there are different forms of microblogs in social networks, we adopt parallel ways to produce the distributed representations of images and texts in the first step so that the model can easily handle different types of inputs. For example, the proposed model can still perform the hashtag recommendation task with only image information when the text is not available or too short and vice versa, making the model more flexible and feasible.

In summary, the AMNN model is presented for hashtag recommendation using encoder–decoder architecture in this article. The main contributions can be summarized as follows.

1) We investigate the task of hashtag recommendation for multimodal microblogs and propose an attention-based neural network framework to extract features from both images and texts and capture correlations between hashtags, called AMNN model.
2) We explore the sequence pattern of hashtags and formulate the hashtag recommendation task as a sequence generation problem. To generate the hashtag sequence, we employ the encoder–decoder architecture to model this generation process.
3) Extensive experimental results on three benchmark data sets, including a large collection, crawled from the Instagram, and two public data sets, demonstrate that our proposed approach yields the best performance for hashtag recommendation by fully exploiting the text and image features. Furthermore, the proposed model also outperforms other baseline methods when there is only the image or text information available.

The rest of the article is organized as follows. Section II introduces the background knowledge and related works. Section III presents the overall design of the proposed AMNN model. Section IV shows the experimental results and evaluates the performance of the model. Finally, Section V concludes this article and highlights our future works.

II. Related Works

Instead of conventional machine learning algorithms, we mainly focus on related works based on neural networks. In this section, we first briefly review the task of hashtag recommendation. Then, we introduce recent works about hashtag recommendations based on neural networks.

A. Traditional Hashtag Recommendation

Compared with the tag recommendation, which is more inclined to describe the website resources, such as music, movie, and document [23], [24], the hashtag recommendation has wider application prospects in social media. Along with the growth of social networks, much effort has been vested in the hashtag recommendation task in recent years. Most of existing studies formulate this task as a classification problem and employ the data mining algorithms, such as topic models [7]–[9], [13], [25], semantic similarity [4], [26], and rank learning [10], [27]. Feng Xiao *et al.* [28] proposed a news-topic oriented hashtag recommendation method using the word co-occurrence pattern, which can help users participate in the discussion of breaking news. Similarly, Godin *et al.* [8] employed topic models to learn the underlying topic assignment of language classified tweets and suggested hashtags to a tweet based on the topic distribution. Zangerle *et al.* [4] presented an approach based on analyzing messages similar to the message the user currently enters. They evaluated multiple similarity measures and presented different ranking techniques for sorting the hashtag candidates. A novel approach for recommending hashtags for tweets is proposed by Li *et al.* [10], which uses learning to the rank algorithm to incorporate features built from topic enhanced word embeddings, tweet entity data, hashtag frequency, hashtag temporal data, and tweet URL domain information. The proposed model generates a hashtag candidate list based on hashtag popularity (frequency) and hashtag temporal information (time).

Besides, there is a popular research line to address the problem of personalized recommendation [5], [29], [30]. Especially, for hashtag recommendation, Wang *et al.* [5] proposed to predict an active user's hashtag usage preference in a collaborative filtering manner and recommend hashtags by relevant scores for a specific microblog. Zhao *et al.* [29] proposed Hashtag-LDA that jointly models the relations between users, hashtags, and words through latent topics and finds the most related hashtags for users by user-distributions and hashtag frequencies. The attention-based neural tag recommendation model (ABNT) [23] utilizes the multilayer perceptron to model the nonlinearities of interactions among users, items, and tags and introduces an attention network to capture the complex pattern of the user's tagging sequence.

Beyond the content and user information, various kinds of external knowledge have been proved to be useful factors for recommendation systems, such as geographical information [31] and temporal information [32], [33]. Location-aware hashtag recommendation model [31] recommends with known GPS locations by learning the relevance of regions with the local popularity of the hashtag. Topic-over-time mixed membership model (TOT-MMM) [32] that introduces the life cycle of the hashtag with popularity recommends hashtags based on the temporal clustering effect of latent topics in tweets. Meanwhile, temporal hashtag reuses have also been analyzed in [33], and a time-dependent and cognitive-inspired hashtag recommendation approach was proposed. Recently, Kou *et al.* [34] proposed a novel hashtag recommendation model based on multifeatures of microblogs, including word embeddings, user-hashtag matrix, and user-hashtag topic model, to alleviate the data sparsity problem. Dey *et al.* [35] proposed the EmTaggeR, which is a word embedding-based method for hashtag recommendation by deriving the word vectors and hashtag embeddings on microblog posts.

However, the abovementioned hashtag recommendation models are not suitable when performing on multimodal microblogs. Thus, more neural-based approaches are proposed and achieve impressive performance; our work is also one of the studies in this field.

B. Neural Network-Based Hashtag Recommendation

Neural network models, especially those involving the attention mechanism, have been largely studied in neuroscience and computer vision [36]. The attention mechanism is originated from the intuition that people select the most pertinent piece of information rather than using all available information [18]. Similarly, the neural network generates better results when focusing on specific parts of the input. Hence, many attention-based neural network methods have been proposed.

Gong and Zhang [14] proposed an attention-based convolutional neural network (CNN) architecture that considers both global and local information to perform the hashtag recommendation task. Huang *et al.* [15] introduced a hierarchical attention mechanism to extend the end-to-end memory network architecture for incorporating the histories of users. Thus, textual information and the corresponding user interests are combined when recommending hashtags. Li *et al.* [13] proposed a topic attention-based long short-term model (LSTM) model that incorporates topic modeling into the LSTM architecture through an attention mechanism. Subsequently, they jointly modeled the content attention and topic attention simultaneously and proposed a novel topical coattention network (TCAN) [16]. However, the hashtag recommendation for images has barely been studied. Recently, Park *et al.* [37] introduced a novel benchmark, titled HAshtag Recommendation for Real-world Images in SOcial Networks (HARRISON), which consists of a visual feature extractor based on the CNN and a multilabel classifier based on neural network. Following this work, Wu *et al.* [38] proposed an attention-based neural image hashtagging network (A-NIH) to model the sequence relationship between social images and hashtags. In this article, we extend this work and transfer the idea to the multimodal social situation.

With respect to the multimodal neural network model, there are also many relevant studies, such as visual question-answering tasks [39]–[41], named entity recognition [42], and mention recommendation [43]. Particularly, Zhang *et al.* [11] proposed a stacked two-layer coattention network for hashtag recommendation task, which incorporates tweet-guided visual attention and image-guided textual attention sequentially. As the coattention network model depends on both textual and visual information, it cannot achieve promising performance when only one type of data is available. Furthermore, a generative method is proposed by Gong *et al.* [12] to recommend hashtags for these multimodal microblogs by incorporating textual and visual information, in which the hashtag suggestion task is converted as a translation problem from visual and textual words to hashtags, with the assumption that hashtags, textual content, and visual words in a microblog describe the same thing using different languages.

From a brief review, it can be observed that only a few studies focused on multimodal information and most of them cannot be directly applied to image-only or text-rarely situations. Meanwhile, few models consider the hashtag correlations. To overcome this problem, we propose a novel AMNN model, which extracts the multimodal features in parallel ways and converts the hashtag recommendation task into a sequence generation problem.

III. Our Model

Unlike most neural network methods, which have treated the hashtag recommendation task as a multiclass classification problem [44], we convert this task as a sequence generation problem in this article. Given a multimodal microblog with texts and images, our model aims to generate a proper hashtag list automatically. The overall architecture of the proposed sequence-to-sequence model is illustrated in Fig. 2, which

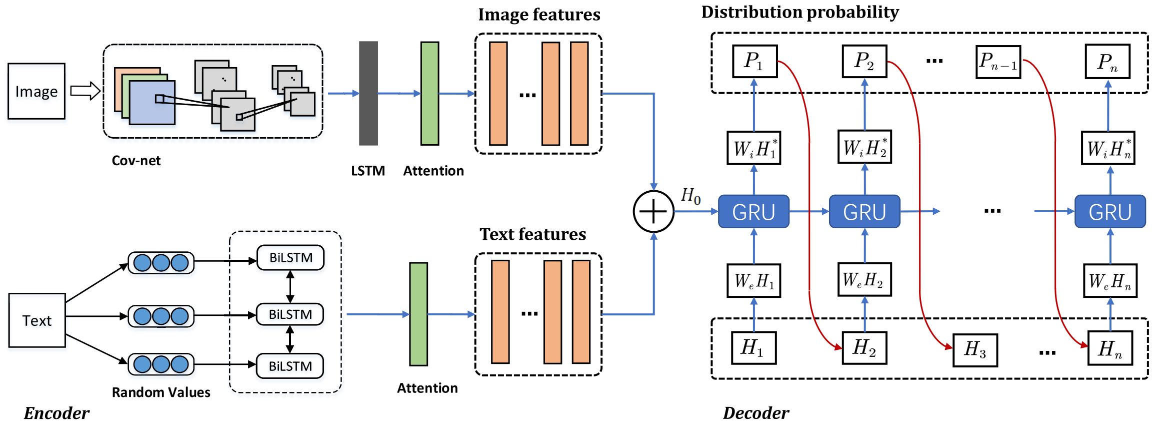


Fig. 2. Overall architecture of the AMNN model.

TABLE I

NOTATIONS IN THIS ARTICLE

Variable	Interpretation
W_* and U_*	weight matrices
b_*	bias vectors
m	maximum number in one hashtag sequence
n	image feature size
d	representation dimension
L	hashtag dictionary size
h_t	hidden state of LSTM at step t
h_t^g	output of GRU at step t
$\overline{h_i}$	i-th immediate hidden state
p_i, p_x	attention probability weight of image and text respectively
X_f	distributed representation of f-th image
T_f	distributed representation of f-th text
P_t	distribution probability of each hashtag at step t
$\odot$	element-wise multiplication
σ	sigmoid function

The subscript * represents a variable symbol.

consists of a hybrid feature extraction encoder and a coupled decoder for the recommendation. In the encoder, image and text features of the microblog are extracted by the parallel neural networks separately and fed into the decoder part after merged. Then, the hashtag sequence is generated based on the hashtag probability that is obtained by the GRU network. Here, we will first introduce the details of feature extraction in the encoder. Then, we present the process in the decoder part when recommending hashtags. Finally, we describe how to train the multimodal neural network model.

For ease presentation of our model, we first define the notations used in this article and list them in Table I.

A. Encoder of Multimodal Neural Network

1) Image Feature Extraction: To extract the image representations, we adopt a hybrid neural network architecture. In the first step, the preliminary feature map of a given image is captured by the CNN. Motivated by the successful use in the field of image processing, we choose the representative neural networks, Inception V3 [45] and ResNet-50 [46], to perform this task. The obtained feature map can be represented as $I_f = [v_1, v_2, \ldots, v_n]$. Then, we employ the LSTM to process the intermediate features sequentially. The transition equations in LSTM are defined as follows:

$$i_t = \sigma(W_i I_f + U_i h_{t-1} + b_i) \tag{1}$$
$$f_t = \sigma(W_f I_f + U_f h_{t-1} + b_f) \tag{2}$$
$$o_t = \sigma(W_o I_f + U_o h_{t-1} + b_o) \tag{3}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c I_f + U_c h_{t-1} + b_c) \tag{4}$$
$$h_t = o_t \odot \tanh(c_t) \tag{5}$$

where $W_* \in \mathbb{R}^{d \times n}$, $U_* \in \mathbb{R}^{d \times d}$, and $b_* \in \mathbb{R}^d$. Other components include input gate i_t, forget gate f_t, output gate o_t, cell state c_t, and hidden state h_t.

To guarantee the consistency of the dimension in the process of processing, the LSTM network is composed of d cells, the same with the representation dimension. Then, the outputs of each cell are formed as a sequence $[h_1, h_2, \ldots, h_d]$ in time order. With the observation that only specific regions of an image are contributed to hashtags, we adopt the attention mechanism to capture spatial features, which contains more information and filters out the noisy data. Hence, we lay an attention layer to generate the image attention distribution. Specifically, we use a single fully connected neural network to project the hidden states into a fixed feature space. Then, a softmax layer is added to output the probability distributions. The calculation process is as follows:

$$\overline{h_i} = \tanh(W_{\overline{v}} h_i + b_{\overline{v}}) \tag{6}$$
$$p_i = \text{softmax}(W_v \overline{h_i} + b_v) \tag{7}$$

where tanh is the activation function.

Finally, the final feature representation of image X_f is represented as $\{X_{f,1}, X_{f,2}, \ldots, X_{f,i}, \ldots, X_{f,d}\}$. Each value in the weighted vector is obtained by (8) with the attention probability, where $X_{f,i}$ is exactly the representation value of the ith dimension

$$X_{f,i} = p_i h_i^{\mathrm{T}}. \tag{8}$$

2) Text Feature Extraction: Considering that the BiLSTM model, which is capable of learning long-term dependencies, has been widely used for NLP tasks in recent years, we employ the BiLSTM to perform the text feature extraction task. BiLSTM obtains the output values from both directions (forward and backward) so that the contextual information can be incorporated well.

For a given text, the representations of each token in the text are randomly initialized with a predefined dimension. Then, the BiLSTM network obtains the forward hidden state $\overrightarrow{h_f}$ and backward hidden state $\overleftarrow{h_b}$. We concatenate both states as the immediate hidden encoding $\overline{h_t}$. After that, we feed $\overline{h_t}$ into the attention layer, and the following process is formulated as follows:

$$\overline{h_t} = \tanh(W_t h_t + b_t) \tag{9}$$

$$p_x = \text{softmax}(\overline{h_t}) \tag{10}$$

$$T_f = p_x \odot h_t. \tag{11}$$

The softmax function is used to get the normalized weights, and T_f is the target output of the attention layer, exactly the text features.

3) Feature Merging: As introduced in Section I, texts and images contain the target information of a social microblog from a different perspective sometimes so that more proper hashtags can be recommended by incorporating both features. To achieve this purpose, we merge the representations of texts and images before recommending, as shown in Fig. 2. The model obtains a comprehensive representation x_{all} of the multimodal microblog by concatenating image and text distributed representations.

It is worth noticing that only image or text features are still enough to perform the hashtag recommendation task in our method, making the model more flexible for most situations in social networks.

B. Decoder for Hashtag Recommendation

The decoder is a gated recurrent unit (GRU) network [47]. In the training process, we first map the hashtag into a fixed feature space as inputs of GRU networks. As a variant of the recurrent neural network (RNN), GRU can well model the correlations between hashtags and multimodal data. The training process is formulated as follows:

$$x_h = W_e H_t, t \in \{1, \ldots, m\} \tag{12}$$

$$h^g_{t+1} = \text{GRU}(x_t), t \in \{1, \ldots, m\} \tag{13}$$

where $H = (H_1, \ldots, H_m)$ is the hashtag sequence. Each hashtag is represented as a one-hot vector H_t with the dimension of hashtag dictionary size L. $W_e \in \mathbb{R}^{d\times m}$ is the weight matrix at time step t, and x_t is the embedding feature of hashtag at time step t.

There is a key difference between the training and prediction process that only multimodal feature maps are used as input for prediction. Consequently, the connections between previous output and successive input of each GRU unit are removed because the hashtag sequence is not available in the prediction process.

Given the previous output h^g_{t-1}, GRU unit uses the update gate z_t and reset gate r_t to generate next output h^g_t. The transition equations are defined as follows:

$$r_t = \sigma\left(W_r x_t + U_r h^g_{t-1}\right) \tag{14}$$

$$z_t = \sigma\left(W_z x_t + U_z h^g_{t-1}\right) \tag{15}$$

$$\overline{h}_t = \tanh\left(W_{\overline{h}} x_t + U_{\overline{h}} r_t \odot h^g_{t-1}\right) \tag{16}$$

$$h^g_t = (1 - z_t) h^g_{t-1} + z_t \overline{h}_t \tag{17}$$

where $W_* \in \mathbb{R}^{d\times l}$ and $U_* \in \mathbb{R}^{d\times d}$.

At each time step, we adopt a single fully connected network and a softmax function to generate the distribution probability of each hashtag at step t

$$P_t = \text{softmax}\left(W_i h^g_s\right) \tag{18}$$

where $h^g = \{h^g_1, h^g_2, \ldots, h^g_m\}$ is the input, and $W_i \in \mathbb{R}^{d\times L}$.

Finally, we can get a ranked list of candidate hashtags and recommend the most related one at time step t, according to the probability of each hashtag.

C. Recommendation

We perform the hashtag recommendation with the following steps. For a given multimodal microblog with a mixture of image and text, the AMNN model first extracts the feature vectors of image and text separately using the hybrid neural network with attention mechanism (encoder). Then, AMNN merges these two representations and feeds the output values into GRU networks to generate a sequence of the recommended hashtags (decoder). The whole process is illustrated in Fig. 2. For the sake of simplicity, we use a greedy search algorithm to generate the final top-K hashtags for the recommendation. As the output words are often repetitive, we also sift out the repeated words at each step.

D. Training

The objective function is the negative log-likelihood of the predicted probability for the ground-truth hashtags at each time step as follows:

$$J = -\sum_{t=0}^{m} \log p(H_t | I, H_0, \ldots, H_{t-1}; \theta) \tag{19}$$

where H_t is the ground-truth hashtag at time step t, H_0 is the output value of encoder, $H_{\{1,t-1\}}$ is the previous hashtag sequence, I is the representation of multimodal microblog, and θ represents all parameters of our proposed network, with respect to all the parameters in encoder and decoder. The training process minimizes the abovementioned loss function using gradient descent. Moreover, the stochastic gradient descent (SGD) method is adopted to update the parameters with the Adam [48] rule. Moreover, L2 and Dropout [49] regularization are used to improve the generalization ability for preventing the model from overfitting.

IV. Experiments

A. Data sets

To evaluate our model, we perform the hashtag recommendation task using two public data sets, HARRISON [37]

TABLE II
SUMMARY STATISTICS FOR THE DATA SETS

DataSet	Categories	Images	Hashtags	Ave	Min	Max
MM-INS	1000	56861	667227	11.73	1	30
HARRISON	997	57383	260000	4.5	2	10
NUS-WIDE	1000	245603	1467922	5.98	1	131

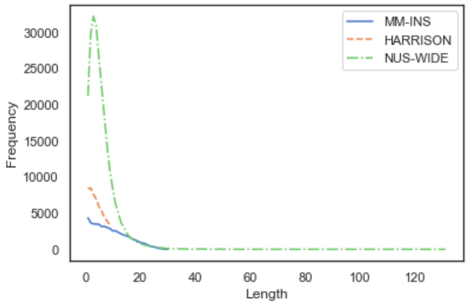


Fig. 3. Comparison curve of sequence length on different data sets.

and NUS-WIDE [50], and a collection of crawled microblogs from Instagram, named MM-INS [2] by using Instaloader API.[3] A brief introduction of these data sets is given as follows.

1) *HARRISON:* A novel benchmark for image hashtag recommendation, titled HARRISON. This data set is composed of 57 383 photographs from Instagram and an average of 4.5 associated hashtags for each photograph.
2) *NUS-WIDE:* A real-world web image data set from the National University of Singapore, titled NUS-WIDE. The raw data set includes 269k images and associated tags from Flickr, with a total number of 5018 unique tags. The data set is cleaned for the usage of experiment comparison that low-frequency tags and samples without tags are dropped out.
3) *MM-INS:* We collect more than 248k public microblogs according to the most 97 popular hashtags that are selected out manually from "Top 100" on Instagram. Hashtags that repeat in one microblog or not in the top-1000 most frequent ones are cleaned. After sifting out microblogs with only image or text, the final collection contains 56 861 samples with both text and image, called MultiModal data from Instagram (MM-INS).

The detailed statistic is shown in Table II. It can be observed that microblogs in MM-INS provide much more hashtags with a higher mean value and a more reasonable range. As shown in Fig. 3, all data sets are imbalanced on sequence length, but the scale on MM-INS is more balanced in the length between 1 and 15. More importantly, individual microblog in HARRISON and NUS-WIDE does not have attached text content. Zhang *et al.* [11] also provided an open tweet data set with hashtags. However, most microblogs in this data set have only 1–3 hashtags, which is not sufficient for a sequence generation task so that we have to abandon it in our experiments.

In the experiment, we adopt the pretrained Inception V3 and ResNet-50 that perform well in ImageNet classification competition [51] to extract the visual features, and all images in experimental data set are rescaled to satisfy the network, 299 × 299 for Inception V3 and 224 × 224 for ResNet-50. Then, we employ the LSTM network to obtain more focused image representations using the attention mechanism. For text words, stop words and low-frequency words are filtered out in preprocessing. Texts are embedded into distributed vectors by the BiLSTM network. We split the data set into a training set and a test set with a ratio of 9:1.

B. Experimental Settings

We use accuracy, precision, and recall as evaluation metrics to measure the overall performance. For comparison with the proposed model, we evaluate the following baselines for hashtag recommendation.

1) *Neighbor Voting:* The neighbor voting (NV) algorithm [52] efficiently learns tag relevance by accumulating votes from visual neighbors and recommends the top-ranked tags accurately. To be fair, we implement this method with image features extracted by the AMNN model.
2) *VGG-Object + VGG-Scene Model:* This baseline algorithm is proposed in HARRISON [37], which consists of two parts. First, VGG-Object and VGG-Scene are used to extract visual features. Then, a multilabel classifier is conducted to recommend the K number of hashtags using concatenation features.
3) *Neural Image Caption:* NIC [17] model is an end-to-end neural network that can automatically generate descriptions for images. NIC is based on a convolution neural network that encodes an image into a compact representation, followed by an RNN that generates a corresponding sentence.
4) *Show, Attend, and Tell:* This baseline model [18] is an attention-based model that automatically learns to describe the content of images. The open-source code is accessible.[4] For a fair comparison, this model is trained with the same label sequence in our experiment.
5) *Multilabel Classification Using ResNet 50:* The hashtag recommendation task can be modeled as a multilabel classification (MLC) task so that we employ the ResNet 50 network [46] to form an end-to-end AMNN+ multilabel baseline to perform this task.
6) *Support Vector Machine:* We also implement a multiclass support vector machine (SVM) classification with LibSVM [53]. The features used are distributed representations of microblogs extracted by the AMNN model.
7) *Multilabel K-Nearest Neighbors:* The multilabel K-nearest neighbors (MLKNN) [54] method uses K-Nearest Neighbors to find the nearest examples to a test class and uses the Bayesian inference to select assigned labels. We employ the MLKNN method to implement the hashtag recommendation task given the text information.

[2] DOI:10.21227/j1rf-fa09

[3] https://instaloader.github.io/

[4] https://github.com/DeepRNN/image_captioning

TABLE III

EVALUATION RESULTS COMPARED WITH BASELINE METHODS ON HARRISON

Model	Precision(%)	Recall(%)	Accuracy(%)
Neighbor Voting	27.04	18.50	44.50
VGG-Object + VGG-Scene	30.02	21.74	52.52
Neural Image Caption	32.32	12.90	35.54
AMNN (Inception V3)	**32.96**	24.08	59.75
AMNN (ResNet 50)	32.36	**25.30**	**60.90**

TABLE IV

EVALUATION RESULTS COMPARED WITH BASELINE METHODS ON NUS-WIDE

Model	Precision(%)	Recall(%)	Accuracy(%)
Neighbor Voting	**32.46**	20.05	54.05
VGG-Object + VGG-Scene	29.27	18.22	55.91
Neural Image Caption	23.50	13.42	44.63
AMNN (Inception V3)	31.00	**20.14**	**59.55**
AMNN (ResNet 50)	29.67	19.3	58.44

TABLE V

EVALUATION RESULTS COMPARED WITH BASELINE METHODS ON MM-INS

Model	Precision(%)	Recall(%)	Accuracy(%)
MLKNN	1.02	0.53	3.43
SVM	14.25	0.11	1.43
Neighbor Voting*	16.46	4.87	32.18
MLC-ResNet 50*	24.83	2.80	29.86
Co-Attention Network	22.55	5.20	3.75
VGG-Object + VGG-Scene*	32.13	12.02	61.32
Neural Image Caption*	27.42	10.12	36.02
Show, Attend and Tell*	30.57	11.08	53.90
AMNN (Inception V3)*	33.08	11.88	61.33
AMNN (ResNet 50)*	34.66	13.01	64.16
AMNN (Inception V3)	34.37	12.75	62.66
AMNN (ResNet 50)	**35.70**	**13.40**	**65.49**

[a] Models which is ended with * only use image features for recommendation.

[b] Only use samples with no less than 5 hashtags when calculating.

8) *Coattention Network:* This model is an integrated framework of visual and textual information for hashtag recommendation tasks using a coattention network, proposed by [11]. The coattention network generates tweet attention and image attention guided by each other.

Although there have been more NIC models recently [18], [19], [21], not all can be well applied for our tasks. That is because the prominent models for image caption always model the strong correlations between words in a text as well as text and images, while the hashtags in one hashtag sequence only have weak correlations, with each other and with the images, which leads to overfitting and generating divergent and repeated outputs.

We implement the proposed model with different CNN networks, Inception V3, and ResNet 50, for image feature extraction in the encoder part. Hashtags follow the original order in the hashtag sequences when training.

C. Experimental Results

1) Overall Evaluation: To evaluate the performance, we compare the proposed model with baseline methods on different data sets. When performing the coattention network [11] on the image-only data set, such as HARRISON and NUS-WIDE, a fixed specific item "PAD" can be set as the missing text so that the coattention network can run on them as well. However, the performance will be greatly reduced. Considering that, we only conduct the coattention network on the MM-INS data set. According to previous works [11], [37], we evaluate the baseline methods with precision, recall, and accuracy as evaluation metrics over different test data sets.

The comparison results with baseline methods on HARRISON and NUS-WIDE are presented in Tables III and IV, respectively. From both tables, we can observe that our model outperforms all other baselines on the two image-only data sets, except the precision of the NV method on NUS-WIDE. As the performance of NV is closely related to the quality of the feature map, the main factor that leads to this result can be attributed to the good use of the AMNN model in feature extraction. Meanwhile, although the performance of NV increases with the growth of data volume, it is noteworthy that the computational overhead also increases exponentially, 10 h on HARRISON and about one week on NUS-WIDE after parallelism in our case. Thus, it cannot be widely applied in social media. Nevertheless, the proposed model gives relative improvement with all metrics; both Inception V3 and ResNet 50 achieve comparable results. This indicates the competitive advantage of the proposed model in hashtag recommendation.

Compared with the NIC model, we can find that the precision results between NIC and our model are close, while the performance gap between recall and accuracy are relatively bigger. This situation illustrates the difference between NIC and our proposed model. NIC model aims to automatically generate texts that describe the content of an image, quite different from the demand of hashtag recommendation and our proposed model. Although the AMNN model describes the content at first, just the same as NIC, the model considers more latent semantic information behind the image to enrich and supplement the meaning during the subsequent sequence generation process. Thus, AMNN can achieve better performance on recall and accuracy.

Then, we compare AMNN with the baseline methods on the MM-INS data set that contains both images and texts. The experimental results are shown in Table V. Observing that AMNN with ResNet 50 performs better than AMNN with Inception V3, AMNN with Inception V3 still outperforms other baseline methods. Although Inception V3 outperforms other architectures in ImageNet competition, we found that the performance of AMNN based on ResNet 50 is more stable and effective for hashtag recommendation. It also shows that the models that incorporate image and text achieve better performance than those using only image features. The results demonstrate that merged features can significantly improve the performance of the hashtag recommendation task. Compared with the coattention network, the proposed models show a much larger performance improvement, about 13% in precision, 8% in recall, and 28% in accuracy. Based on its own characteristics, the coattention network model is more sensitive to the amount of text information, leading to the performance decline when texts are not as rich as on twitter. Although both are social networks, Twitter is a short-form news and social media site that basic format is short text microblogs, while Instagram aims primarily at sharing media.[5] Thus, text-guided visual attention in coattention network may not contribute much for training on Instagram data. On the other hand, the results that our proposed model achieves better performance only rely on image features than other baselines, which, further, shows the advantage of the AMNN model. Since the AMNN model decouples the feature extraction process of different kinds of data, this model can be well applied for different situations, such as image-only (HARRISON and NUS-WIDE) and multimodal (MM-INS).

Compared with ResNet 50 (MLC), the better performance of AMNN demonstrates the effectiveness of sequence generation prediction for hashtag recommendation. Resnet 50 (MLC) model is based on the CNN variant, which transforms the multilabel classification into binary classification problems for each label by adopting the sigmoid cross-entropy loss. Thus, the model will be completely biased with the imbalanced data and impact the final results. According to the statistic information, it is clear that the MM-INS has the maximum average value of associated hashtags for each image and a more reasonable range. The main factors that impact the final result lay on the different label distribution of each data set. As we prefer to download the microblogs with multiple hashtags when crawling, MM-INS has more hashtag information and more uniform frequency distribution. Different from MM-INS, the other two data sets are extremely imbalanced and lead to unacceptable low results. Hence, we only display the result of Resnet 50 (MLC) baseline method on MM-INS for comparison in this article.

It can be observed that the coattention network with textual and visual information performs not good enough and is exceeded by the VGG-Object + VGG-Scene in Table V, which only uses the visual information. As the coattention network baseline model uses multiclass classification to conduct the hashtag recommendation task that adopts the softmax classifier with cross-entropy loss, it is quite different from the prediction problem that we formulated in this article. With the hashtag sequence existing, the recommendation task becomes a multilabel classification problem. Although multilabel classification is a generalization of multiclass classification, which can be treated as a single-label problem that categorizes instances into precisely one of more than two classes, there is no constraint on how many of the classes the instance can be assigned to in the multilabel problem.[6] It explains why the multilabel classification model, such as VGG-Object + VGG-Scene, can perform better, where each microblog has multiple hashtags. As the coattention network model can achieve the ideal performance when data samples have only a single or small number of hashtags, it becomes a litter inferior in this situation where each microblog has multiple hashtags. Moreover, since the baseline coattention network model treats the text as a main part of the hashtag recommendation task for tweets, the sparseness and noise of texts on Instagram and the imbalanced class/hashtag distribution further degrade the performance of the coattention model.

TABLE VI
STATISTICS OF THE TEXT DATA IN MM-INS

Total words	Tokens	Min-len	Max-len	Ave
684,396	47,169	1	248	12.04

Fig. 4 shows the accuracy, precision, recall, and F1-score curves of models with a different number of recommended hashtags on the MM-INS data set. As “Show, Attend, and Tell” is not aimed for hashtag recommendation but sequence generation, most of the output is less than five terms. We have to filter out about 90% of samples when calculating performance so that we do not add this model in Fig. 4 for a fair comparison. Each point of a curve represents the extraction of a different number of hashtags, ranging from 1 to 5. Clearly, the curves of the AMNN model always are the highest in all metrics compared with other models, indicating that the AMNN model significantly improves the performance even though the hashtag number changing. Although the precision value decreases with the number of hashtags increasing, the AMNN model still outperforms the other methods. The precision of ResNet 50 (MLC) descends quickly along with the number of hashtags increasing, as shown in Fig. 4(b). The coattention network baseline also shows a similar curve in precision value, while our proposed model descends more slowly. Moreover, the gaps in accuracy, recall, and F1-score curves are widening when recommended hashtags increasing because correlations emerge easier with longer hashtag sequence so that the AMNN has more advantages. Based on the experimental results, we can conclude that our proposed model and variants outperform all other methods on this data set.

In addition, it should be pointed out that the NV, which achieves good performance on HARRISON and NUS-WIDE data sets, does not perform well in the MM-INS. As microblogs in MM-INS contain both image and text information, the degradation in performance indicates that conventional machine learning algorithms cannot make the best use of multiple types of features. Other algorithms, such as MLKNN and Multilabel classification SVM, also cannot achieve the expected performance. These results show that neural network-based methods are more suitable for the hashtag recommendation task with multimodal demands.

2) Ablation Experiments: Because of the short and sparse characteristics, as shown in Table VI, texts on Instagram may not contain determinant information of the hashtags. Thus, the effectiveness of text features is limited. To evaluate the performance of text features extracted by the BiLSTM network, we experiment with the AMNN model using both text and image information, while text features are replaced with other kinds of distributed representations. Image features are extracted by ResNet 50 for a fair comparison. Table VII shows the results with use of glove,[7] word2vec,[8] and fasttext.[9]

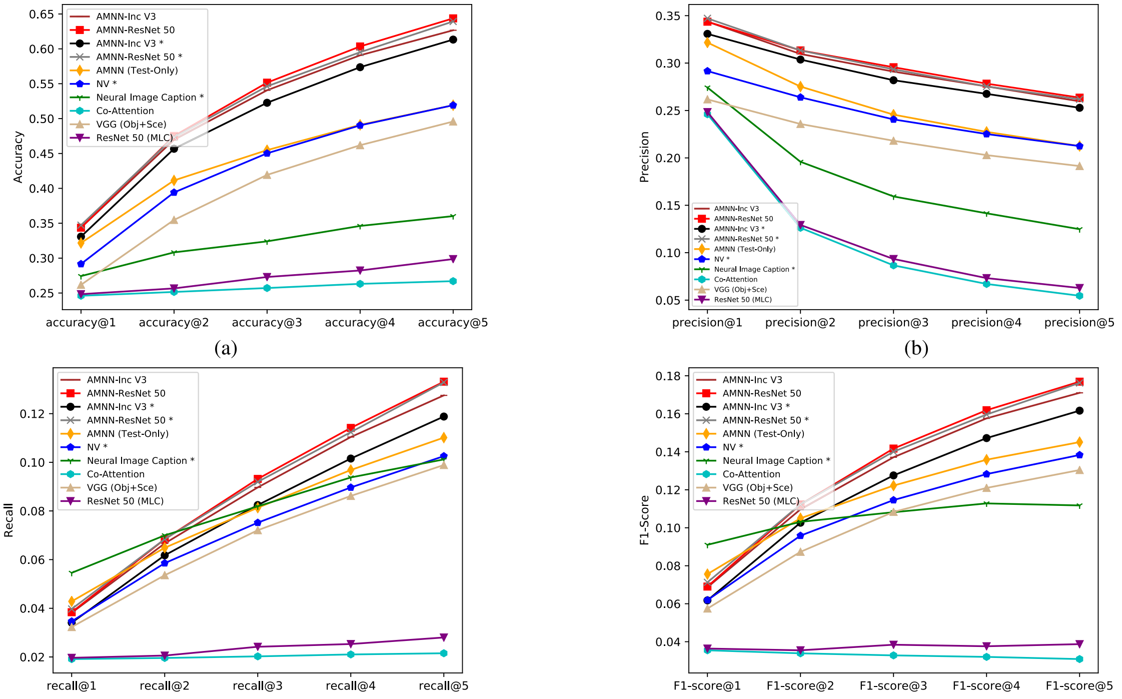
[5] http://www.differencebetween.net/technology/differencesbetween-instagramandtwitter/

[6] https://en.wikipedia.org/wiki/Multilabel_classification

[7] https://nlp.stanford.edu/projects/glove/

[8] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

[9] http://mxnet.incubator.apache.org/api/python/contrib/text.html



(a) (b)

(c) (d)

Fig. 4. (a) Accuracy, (b) precision, (c) recall, and (d) F1-score curves with a different number of recommendation hashtags compared with baseline methods on MM-INS. Note that models ended with "*" only use image features, including the VGG (Obj+Sce) and ResNet 50 (MLC) and part of AMNN models. AMNN (Text-Only) model only uses text features, and other models use both image and text features. Results of SVM and MLKNN are excluded for the low values, "Show, attend, and tell" model for most output length is less than 5.

TABLE VII

RECOMMENDATION RESULTS WHEN REPLACING TEXT FEATURES

Model	Precision (%)	Recall (%)	Accuracy (%)
AMNN + glove	33.35	11.80	60.80
AMNN + word2vec	33.86	12.34	61.50
AMNN + fasttext	34.33	11.86	62.77

TABLE VIII

PERFORMANCE OF AMNN (INCEPTION V3) WHEN RANDOMLY SHUFFLE THE LABEL SEQUENCE

Dataset	Precision (%)	Recall (%)	Accuracy (%)
HARRISON	29.55 (↓3.41)	20.69 (↓3.39)	57.57 (↓2.18)
NUS-WIDE	22.82 (↓8.18)	13.41 (↓6.73)	48.28 (↓11.27)
MM-INS	24.72 (↓9.65)	6.38 (↓6.37)	46.20 (↓16.46)

We use the pretrained word embeddings in a fixed situation as text features for comparison, which means that the learning process of text features is no need to be executed in the variants. Compared with results in Table IV, it shows that the original AMNN model performs better than other variants, indicating that text features learned by the AMNN are feasible for hashtag recommendation. Although the gain is not big because of the limitation of text sparseness, better performance still suggests that AMNN architecture is more effective.

In addition, in order to explore the impact of the hashtag order, we randomly shuffle the label sequence to perturb the original order. As shown in Tables VIII and IX, the performance of AMNN has been affected when the label sequence is randomly shuffled, and the percentage of change is also listed, but the extent and effect of this impact are not the same on different data sets. In most cases, the shuffle operation leads to performance decline. Particularly, the decline is more significant on MM-INS, while the performance on HARRISON also degraded but not so great. The MM-INS has the highest average associated hashtags and contains more long sequences so that the original order on MM-INS can better show the progressive correlation in latent semantics of hashtags. Therefore, the performance of this data set is more susceptible to the order information. This explains the reason why the sequence order has so much impact on the performance of these models. However, the performance on NUS-WIDE is not as expected, seriously polarized when using different neural networks. As the NUS-WIDE data set has a larger cardinality and the distribution in labels is nonuniform, it is harder to capture the correlation between labels and predict the right ones for individual samples. Thus, the performance of the AMNN model on NUS-WIDE becomes unpredictable when hashtags randomly shuffled. Overall, the ablation experimental results show that the order of labels is very important for the model's performance.

TABLE IX

PERFORMANCE OF AMNN (RESNET 50) WHEN RANDOMLY SHUFFLE THE LABEL SEQUENCE

Dataset	Precision (%)	Recall (%)	Accuracy (%)
HARRISON	30.21 (↓2.15)	22.39 (↓2.91)	60.13 (↓0.77)
NUS-WIDE	31.38 (↑1.71)	19.31 (↑0.01)	59.36 (↑0.92)
MM-INS	26.52 (↓9.18)	9.04 (↓4.36)	50.41 (↓15.08)

D. Discussion

In the proposed model, we formulate the hashtag recommendation task as a sequence generation problem with the assumption that correlations exist between hashtags. Under this assumption, we adopt a sequence-to-sequence model to generate the recommended hashtag list. In the encoder, we divide the feature extraction into different parts according to the data modal. Then, multiple features are merged into a unified representation for follow-up recommendations. Integration makes the representation of multimodal microblog more accurate and focused. The decoder is built by GRU networks to generate a hashtag sequence for the recommendation, which can easily capture the correlations between hashtags as a variant of cyclic neural networks.

The better experimental results demonstrate that correlations between hashtags for one microblog do exist and a sequence generation model can achieve comparable results based on this characteristic. Compared with the baseline methods, the proposed model achieves competitive performance using only one type of features, image or text so that the proposed model is powerful for hashtags recommendation in a real and complex social network. Given the abovementioned discussion, it can be concluded that the proposed AMNN model performs better than the state-of-the-art methods in the hashtag recommendation task on multimodal microblogs.

V. Conclusion

In this article, we propose an AMNN model for hashtag recommendation. We convert this task into a sequence generation problem based on the assumption that there exist correlations between hashtags. Particularly, we introduce a hybrid neural network to extract features of multimodal microblogs and adopt a sequence-to-sequence architecture for hashtag recommendation. Under this framework, image and text features can be extracted separately. We evaluate our model against several baseline models in two different situations that recommend hashtags with image-only and mixture of image and text. For the situation of image-only, we found that the AMNN model can significantly improve the performance, and the encoder–decoder architecture is effective for the hashtag recommendation task. We also test the proposed model on a large social data set collected from Instagram and compare it with the state-of-the-art baseline models. The improvement in recommendation performance shows the effectiveness of our model. Furthermore, the proposed AMNN model also achieves relative performance improvement when only image or text information exists. The overall experimental results demonstrate that our model is capable of achieving better performance for hashtag recommendation with multimodal microblogs.

There are a few directions that we would like to investigate in the future. We want to consider more external knowledge, such as comments on each microblog and user information. We will further collect this kind of information in the future and provide a domain-specific and personalized recommendation. Furthermore, we also want to investigate the potential ways for the improvement of our model based on recent competitive architectures on image captioning. The significant achievements of the nonautoregressive method in neural machine translation also motivate us. It is a challenging task to explore the way to incorporate this method with the characteristics of hashtag prediction. We left these issues to be further optimized and validated in our future work.

References

[1] M. Efron, "Hashtag retrieval in a microblogging environment," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Geneva, Switzerland, Jul. 2010, pp. 787–788, doi: 10.1145/1835449.1835616.

[2] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using Twitter hashtags and smileys," in *Proc. 23rd Int. Conf. Comput. Linguistics*, Beijing, China, Aug. 2010, pp. 241–249. [Online]. Available: http://aclweb.org/anthology/C/C10/C10-2028.pdf

[3] Z. Ma, A. Sun, and G. Cong, "On predicting the popularity of newly emerging hashtags in Twitter," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 7, pp. 1399–1410, Jul. 2013, doi: 10.1002/asi.22844.

[4] E. Zangerle, W. Gassler, and G. Specht, "On the impact of text similarity functions on hashtag recommendations in microblogging environments," *Social Netw. Anal. Mining*, vol. 3, no. 4, pp. 889–898, Dec. 2013, doi: 10.1007/s13278-013-0108-x.

[5] Y. Wang, J. Qu, J. Liu, J. Chen, and Y. Huang, "What to tag your microblog: Hashtag recommendation based on topic analysis and collaborative filtering," in *Proc. 16th Asia–Pacific Web Conf. (APWeb)*, Changsha, China, Sep. 2014, pp. 610–618, doi: 10.1007/978-3-319-11116-2_58.

[6] B. Shi, G. Ifrim, and N. Hurley, "Learning-to-rank for real-time high-precision hashtag recommendation for streaming news," in *Proc. 25th Int. Conf. World Wide Web (WWW)*, Montreal, QC, Canada, Apr. 2016, pp. 1191–1202, doi: 10.1145/2872427.2882982.

[7] Z. Ding, Q. Zhang, and X. Huang, "Automatic hashtag recommendation for microblogs using topic-specific translation model," in *Proc. 24th Int. Conf. Comput. Linguistics (COLING)*, Mumbai, India, Dec. 2012, pp. 265–274. [Online]. Available: http://aclweb.org/anthology/C/C12/C12-2027.pdf

[8] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle, "Using topic models for Twitter hashtag recommendation," in *Proc. 22nd Int. Conf. World Wide Web (WWW Companion)*, Rio de Janeiro, Brazil, May 2013, pp. 593–596, doi: 10.1145/2487788.2488002.

[9] J. She and L. Chen, "TOMOHA: Topic model-based hashtag recommendation on Twitter," in *Proc. 23rd Int. Conf. World Wide Web (WWW Companion)*, Seoul, South Korea, Apr. 2014, pp. 371–372, doi: 10.1145/2567948.2577292.

[10] Q. Li, S. Shah, A. Nourbakhsh, X. Liu, and R. Fang, "Hashtag recommendation based on topic enhanced embedding, tweet entity data and learning to rank," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Indianapolis, IN, USA, Oct. 2016, pp. 2085–2088, doi: 10.1145/2983323.2983915.

[11] Q. Zhang, J. Wang, H. Huang, X. Huang, and Y. Gong, "Hashtag recommendation for multimodal microblog using co-attention network," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 3420–3426, doi: 10.24963/ijcai.2017/478.

[12] Y. Gong, Q. Zhang, and X. Huang, "Hashtag recommendation for multimodal microblog posts," *Neurocomputing*, vol. 272, pp. 170–177, Jan. 2018, doi: 10.1016/j.neucom.2017.06.056.

[13] Y. Li, T. Liu, J. Jiang, and L. Zhang, "Hashtag recommendation with topical attention-based LSTM," in *Proc. 26th Int. Conf. Comput. Linguistics (COLING)*, Osaka, Japan, Dec. 2016, pp. 3019–3029. [Online]. Available: http://aclweb.org/anthology/C/C16/C16-1284.pdf

[14] Y. Gong and Q. Zhang, "Hashtag recommendation using attention-based convolutional neural network," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, New York, NY, USA, Jul. 2016, pp. 2782–2788. [Online]. Available: http://www.ijcai.org/Abstract/16/395

[15] H. Huang, Q. Zhang, Y. Gong, and X. Huang, "Hashtag recommendation using end-to-end memory networks with hierarchical attention," in *Proc. 26th Int. Conf. Comput. Linguistics (COLING)*, Osaka, Japan, Dec. 2016, pp. 943–952. [Online]. Available: http://aclweb.org/anthology/C/C16/C16-1090.pdf

[16] Y. Li, T. Liu, J. Hu, and J. Jiang, "Topical co-attention networks for hashtag recommendation on microblogs," *Neurocomputing*, vol. 331, pp. 356–365, Feb. 2019, doi: 10.1016/j.neucom.2018.11.057.

[17] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3156–3164, doi: 10.1109/CVPR.2015.7298935.

[18] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015, pp. 2048–2057. [Online]. Available: http://jmlr.org/proceedings/papers/v37/xuc15.html

[19] W. Jiang, L. Ma, Y. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proc. 15th Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 510–526, doi: 10.1007/978-3-030-01216-8_31

[20] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1179–1195, doi: 10.1109/CVPR.2017.131.

[21] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7219–7228.

[22] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 711–727, doi: 10.1007/978-3-030-01264-9_42.

[23] J. Yuan, Y. Jin, W. Liu, and X. Wang, "Attention-based neural tag recommendation," in *Proc. 24th Int. Conf. Database Syst. Adv. Appl. (DASFAA)*, Chiang Mai, Thailand, Apr. 2019, pp. 350–365, doi: 10.1007/978-3-030-18579-4_21.

[24] X. Shi, H. Huang, S. Zhao, P. Jian, and Y. Tang, "Tag recommendation by word-level tag sequence modeling," in *Proc. Int. Conf. Database Syst. Adv. Appl. (DASFAA)*, Chiang Mai, Thailand, Apr. 2019, pp. 420–424, doi: 10.1007/978-3-030-18590-9_58.

[25] Y. Gong, Q. Zhang, and X. Huang, "Hashtag recommendation using Dirichlet process mixture models incorporating types of hashtags," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, Sep. 2015, pp. 401–410. [Online]. Available: http://aclweb.org/anthology/D/D15/D15-1046.pdf

[26] N. Ben-Lhachemi and E. H. Nfaoui, "An extended spreading activation technique for hashtag recommendation in microblogging platforms," in *Proc. 7th Int. Conf. Web Intell., Mining Semantics (WIMS)*, Amantea, Italy, Jun. 2017, pp. 16:1–16:8, doi: 10.1145/3102254.3102283.

[27] S. Sedhai and A. Sun, "Hashtag recommendation for hyperlinked tweets," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Gold Coast, QLD, Australia, Jul. 2014, pp. 831–834, doi: 10.1145/2600428.2609452.

[28] F. Xiao, T. Noro, and T. Tokuda, "News-topic oriented hashtag recommendation in Twitter based on characteristic co-occurrence word detection," in *Proc. 12th Int. Conf. Web Eng. (ICWE)*, Berlin, Germany, Jul. 2012, pp. 16–30, doi: 10.1007/978-3-642-31753-8_2.

[29] F. Zhao, Y. Zhu, H. Jin, and L. T. Yang, "A personalized hashtag recommendation approach using LDA-based topic model in microblog environment," *Future Gener. Comput. Syst.*, vol. 65, pp. 196–206, Dec. 2016, doi: 10.1016/j.future.2015.10.012.

[30] J. Niu, L. Wang, X. Liu, and S. Yu, "FUIR: Fusing user and item information to deal with data sparsity by using side information in recommendation systems," *J. Netw. Comput. Appl.*, vol. 70, pp. 41–50, Jul. 2016, doi: 10.1016/j.jnca.2016.05.006.

[31] R. Pálovics, P. Szalai, L. Kocsis, J. Pap, E. Frigó, and A. A. Benczúr, "Location-aware online learning for top-k hashtag recommendation," in *Proc. Workshop Location-Aware Rec. (LocalRec), 9th ACM Conf. Rec. Syst. (RecSys)*, Vienna, Austria, Sep. 2015, pp. 36–39. [Online]. Available: http://ceur-ws.org/Vol-1405/paper-06.pdf

[32] H.-M. Lu and C.-H. Lee, "A Twitter hashtag recommendation model that accommodates for temporal clustering effects," *IEEE Intell. Syst.*, vol. 30, no. 3, pp. 18–25, May 2015, doi: 10.1109/MIS.2015.20.

[33] D. Kowald, S. C. Pujari, and E. Lex, "Temporal effects on hashtag reuse in Twitter: A cognitive-inspired hashtag recommendation approach," in *Proc. 26th Int. Conf. World Wide Web*, Perth, WA, Australia, Apr. 2017, pp. 1401–1410, doi: 10.1145/3038912.3052605.

[34] F.-F. Kou *et al.*, "Hashtag recommendation based on multi-features of microblogs," *J. Comput. Sci. Technol.*, vol. 33, no. 4, pp. 711–726, Jul. 2018, doi: 10.1007/s11390-018-1851-2.

[35] K. Dey, R. Shrivastava, S. Kaushik, and L. V. Subramaniam, "EmTaggeR: A word embedding based novel method for hashtag recommendation on Twitter," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, New Orleans, LA, USA, Nov. 2017, pp. 1025–1032, doi: 10.1109/ICDMW.2017.145.

[36] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015, doi: 10.1109/TMM.2015.2477044.

[37] M. Park, H. Li, and J. Kim, "HARRISON: A benchmark on hashtag recommendation for real-world images in social networks," 2016, *arXiv:1605.05054*. [Online]. Available: https://arxiv.org/abs/1605.05054

[38] G. Wu, Y. Li, W. Yan, R. Li, X. Gu, and Q. Yang, "Hashtag recommendation with attention-based neural image hashtagging network," in *Proc. 25th Int. Conf. Neural Inf. Process. (ICONIP)*, Siem Reap, Cambodia, Dec. 2018, pp. 52–63, doi: 10.1007/978-3-030-04179-3_5.

[39] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 2016, pp. 289–297. [Online]. Available: http://papers.nips.cc/paper/6202-hierarchical-question-image-co-attention-for-visual-question-answering

[40] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1839–1848, doi: 10.1109/ICCV.2017.202.

[41] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2156–2164, doi: 10.1109/CVPR.2017.232.

[42] 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. (NAACL-HLT)*, New Orleans, LA, USA, Jun. 2018, pp. 852–860. [Online]. Available: https://www.aclweb.org/anthology/N18-1078/

[43] H. Huang, Q. Zhang, and X. Huang, "Mention recommendation for Twitter with end-to-end memory network," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Melbourne, VIC, Australia, Aug. 2017, pp. 1872–1878, doi: 10.24963/ijcai.2017/260.

[44] R. Ma, X. Qiu, Q. Zhang, X. Hu, Y.-G. Jiang, and X. Huang, "Co-attention memory network for multimodal microblog's hashtag recommendation," *IEEE Trans. Knowl. Data Eng.*, early access, Aug. 1, 2019, doi: 10.1109/TKDE.2019.2932406.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[47] K. Cho, B. van Merrienboer, C. C. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1724–1734. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1179.pdf

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–15, [Online]. Available: http://arxiv.org/abs/1412.6980

[49] N. Srivastava, G. E. 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. [Online]. Available: http://dl.acm.org/citation.cfm?id=2670313

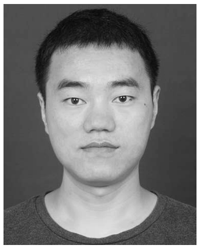
[50] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Conf. Image Video Retr. (CIVR)*, Santorini, Greece, Jul. 2009, pp. 1–9.

[51] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[52] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

[53] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

[54] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007.



Qi Yang received the B.S. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2014, where he is currently pursuing the Ph.D. degree.

His research interests include big data analysis and data mining.

Gaosheng Wu received the B.S. degree in computer science from Central China Normal University, Wuhan, China, in 2016, and the M.S. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2019, where he is currently pursuing the master's degree with the School of Computer Science and Technology.

His research interest includes deep learning and recommendation systems.

Yuhua Li (Member, IEEE) received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2006.

She is currently an Associate Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology. Her research interests include data mining, machine learning, and big data analysis.

Dr. Li is also a member of the ACM and a Senior Member of the China Computer Federation (CCF).

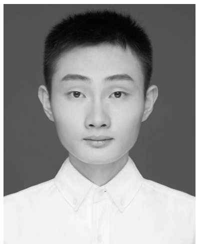
Ruixuan Li (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 1997, 2000, and 2004, respectively.

He was a Visiting Researcher with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, from 2009 to 2010. He is currently a Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology. His research interests include cloud and edge computing, big data management, and distributed system security.

Dr. Li is also a member of ACM.

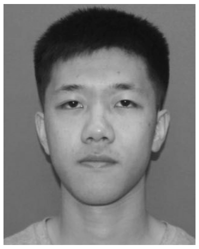
Xiwu Gu received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2007.

He is currently an Associate Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology. His research interests include distributed systems, big data, and middleware.



Huicai Deng received the B.S. degree from the College of Computer Science and Technology, Chongqing University, Chongqing, China, in 2019. He is currently pursuing the master's degree with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China.

His research interests include machine learning and network embedding.



Junzhuang Wu received the B.S. degree in information science and engineering from Hunan University, Changsha, China, in 2019. He is currently pursuing the master's degree with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China.

His research interests include data mining and artificial intelligence.