Hashtag Recommendation for Photo Sharing Services

Suwei Zhang[1], Yuan Yao[1], Feng Xu[1], Hanghang Tong[2], Xiaohui Yan[3], Jian Lu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Arizona State University, USA [3]Poisson Lab, Huawei Technologies, China
zsw@smail.nju.edu.cn, {y.yao, xf, lj}@nju.edu.cn, hanghang.tong@asu.edu, yanxiaohui2@huawei.com

Abstract

Hashtags can greatly facilitate content navigation and improve user engagement in social media. Meaningful as it might be, recommending hashtags for photo sharing services such as Instagram and Pinterest remains a daunting task due to the following two reasons. On the *endogenous* side, posts in photo sharing services often contain both images and text, which are likely to be correlated with each other. Therefore, it is crucial to coherently model both image and text as well as the interaction between them. On the *exogenous* side, hashtags are generated by users and different users might come up with different tags for similar posts, due to their different preference and/or community effect. Therefore, it is highly desirable to characterize the users' tagging habits. In this paper, we propose an integral and effective hashtag recommendation approach for photo sharing services. In particular, the proposed approach considers both the endogenous and exogenous effects by a content modeling module and a habit modeling module, respectively. For the content modeling module, we adopt the parallel co-attention mechanism to coherently model both image and text as well as the interaction between them; for the habit modeling module, we introduce an external memory unit to characterize the historical tagging habit of each user. The overall hashtag recommendations are generated on the basis of both the post features from the content modeling module and the habit influences from the habit modeling module. We evaluate the proposed approach on real Instagram data. The experimental results demonstrate that the proposed approach significantly outperforms the state-of-the-art methods in terms of recommendation accuracy, and that both content modeling and habit modeling contribute significantly to the overall recommendation accuracy.

Introduction

Photo sharing services such as Instagram and Pinterest are gaining increasing popularity in recent years. According to the latest statistics in 2018, Pinterest has reached 175 million monthly active users,[1] while the number for Instagram is one billion.[2] Typically, photo sharing services provide users a platform where they can upload their photos with textual descriptions and share them with the public or the pre-approved friends. Optionally, users can attach several hashtags with their uploaded photos. The hastags allow users to indicate the topics of their uploaded content (a.k.a. posts), and they have been shown to be beneficial for many tasks including sentiment analysis (Hasan, Agu, and Rundensteiner 2014), information retrieval (Highfield and Leaver 2015), and topic extraction (Lim and Buntine 2014). Hashtags can also facilitate users to interact with others. For example, it has been shown that posts with at least one hashtag receive 12.6% more user engagement than those without any hashtags on Instagram.[3]

To date, many hashtag recommendation methods have been proposed for online posts (Krestel, Fankhauser, and Nejdl 2009; Sedhai and Sun 2014; Wang et al. 2016). However, an effective hashtag recommendation method for photo sharing services is still in need due to the following two reasons. The first reason lies in the endogenous side: while most of the existing efforts focus on recommending hashtags for either text (Wu et al. 2016; Krestel, Fankhauser, and Nejdl 2009; Li et al. 2016) or images (Wang et al. 2016; Gong et al. 2014; Wei et al. 2014), the post in photo sharing services usually consists of both image and text, which are likely to be correlated with each other. Therefore, a coherent modeling of image and text as well as the interaction between the two is crucial to better understanding the semantics of the post, which in turn affects hashtag recommendation performance. The second reason is about the exogenous side: different users may have different tagging habits by attaching different hashtags to similar posts, and such tagging habits can be attributed to many aspects including personal preference or community effect (Zhu, Aloufi, and El Saddik 2015). Although some personalized treatments exist for the problem (e.g., tensor factorization (Fang et al. 2015; Nguyen, Wistuba, and Schmidt-Thieme 2017) and graph-based methods (Feng and Wang 2012; Guan et al. 2009)), they generate the hashtags mainly based on an implicit collaborative modeling of the existing interactions among users, posts, and hastags. An explicit modeling of user habits in conjunction with content modeling has largely remained absent, especially in the context of photo sharing services.

An illustrative example from Instagram is shown in Table 1, where two similar posts are retrieved from two

[1]https://www.socialpilot.co/blog/social-media-statistics/

[2]https://techcrunch.com/2018/06/20/instagram-1-billion-users/

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

	***Text***: Happy Friday and National Watermelon Day from Paco!! ***Hashtags***: #Friday #paco #nationalwatermelonday #watermelon #bestlabradors #dogsofinstagram #sendadogphoto
	***Text***: Happy National Watermelon Day! Share a watermelon with your dogs to stay cool! ***Hashtags***: #hungrydog #nationalwatermelonday #dogsofinstagram #smile #fundogs #summerlovin

Table 1: An illustrative example from Instagram. This example shows that (1) both image and text are useful for hashtag recommendation, and (2) users may have their tagging habits by attaching different hashtags for similar posts.

different users. First, we can observe from the example that the two users have some common hashtags (i.e., #nationalwatermelonday and #dogsofinstagram) which are related to both the image and the text. Moreover, the first user also attaches #Friday and #paco which can only be inferred from the text. Second, the second user attaches #smile and #summerlovin, indicating his/her distinctive tagging habit. In a nutshell, both content (image and text) and user habits play important roles for effective hashtag recommendation in photo sharing services.

In this paper, we aim to address the limitations of existing work by designing an integral and effective hashtag recommendation approach for the photo sharing services. In particular, we propose a memory augmented co-attention model MACON, which (1) simultaneously models image and text with a content modeling module and (2) introduces external memory to explicitly learn the users' tagging habits with a habit modeling module. To be specific, in the content modeling module, we first extract content features from both image and text with neural networks. Next, since regions in an image or words in a piece of text are not always equally important, we adopt a parallel co-attention mechanism (Lu et al. 2016) so that the image features and text features can guide each other towards the prediction target. For the habit modeling module, inspired by memory networks (Sukhbaatar et al. 2015), we first sample a set of historical posts for each user in the memory unit, and then learn the user habit by connecting the historical posts with the current post and characterizing how the user assigns hashtags to the historical posts. The overall recommendations are generated based on the feature vectors extracted from the content modeling module and the influence vectors from the habit modeling module. To demonstrate the effectiveness of the proposed approach, we perform empirical experiments on a large Instagram dataset. The experimental results show that the proposed method significantly outperforms the state-of-the-art methods in terms of recommendation accuracy.

In summary, the main contributions of this paper include:

- A novel neural network based approach, MACON, to recommend hashtags for photo sharing services. MACON takes both post content modeling and user habit modeling into consideration. To the best of our knowledge, we are the first to integrate both image and text modeling as well as the user habit influence into a single model for hashtag recommendation.
- Experimental results showing significant performance improvements. Compared with the state-of-the-art, the proposed approach can achieve up to 23.6%, 20.8%, and 13.4% improvements in terms of precision, recall, and F1-score, respectively.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed approach, and Section 4 presents the experimental evaluations. Section 5 concludes.

Related Work

In this section, we briefly review the related hashtag recommendation work including content-based recommendations and personalized recommendations. Content-based hashtag recommendations can be further divided into three categories based on the input: image only, text only, and multimodal data.

Hashtag recommendation for images. Traditional hashtag recommendation methods for images mainly focus on the analysis of users' tagging behaviors. For example, Liu et al. (2009) propose a probabilistic ranking method for tagging Flickr photos; Sigurbjörnsson et al. (2008) study the statistical tagging patterns on Flickr and then apply the patterns to recommend tags. Recent methods mainly use CNNs to learn the semantics of images. They either use the features learned from existing CNNs (Wei et al. 2014; Gong et al. 2013), or build an end-to-end model to collectively learn the semantics of images and tags (Gong et al. 2014; Wang et al. 2016).

Hashtag recommendation for text. Recommending hashtags for textual content can be formulated as a document classification problem, which is one of the classical tasks in natural language processing. Most of the existing methods are built upon topic models and neural networks. For example, Wu et al. (2016) introduce a supervised variant of LDA, Weston et al. (2014) adopt CNNs to predict the tags, Gong and Zhang (2016) further integrate an attention mechanism into CNNs, Wang et al. (2015) use autoencoders to model the textual content, Huang et al. (2016) adopt the memory networks, and Li et al. (2016) combine topical distributions with RNNs.

Hashtag recommendation based on multimodal data. Using multiple types of input for hashtag recommendation

has also been studied. For example, Rawat and Kankanhalli (2016) combine image features extracted from CNNs with contextual features such as time and location; Hwang and Grauman (2012) and Zhang et al. (2017) recommend hashtags based on both text and image; Zhu et al. (2015) combine images and users' social clues to make recommendations. The proposed method in this paper also uses the multimodal input of text and image. Different from the existing work, we further model the users' tagging habits.

Personalized hashtag recommendation. Our work is also related to personalized hashtag recommendation. While content-based methods aim to recommend hashtags solely based on the content, personalized methods take the users' historical taggings into account to implicitly model their personal preferences. For example, Qian et al. (2013) construct a tag vocabulary for each user from his/her tagging history; Guan et al. (2009) as well as Feng and Wang (2012) treat the problem as an edge prediction problem in a heterogeneous network; Rendle and Schmidt-Thieme (2010), Fang et al. (2015), and Nguyen et al. (2017) apply tensor factorization on the user-item-tag tensors. Essentially, the above methods are built upon the interactions among users, items, and tags, while the content information is widely ignored. In contrast, we aim to explicitly model the users' habits and simultaneously model multimodal content with user habits.

Others. By treating the historical posts as memory and the current post to tag as query post, our work can be seen as a variant of the memory networks (Weston, Chopra, and Bordes 2014; Sukhbaatar et al. 2015). Our work is also related to visual question answering (Lu et al. 2016) and image captioning (Xu et al. 2015), where both text and image are modeled.

The Proposed Approach

In this section, we present the proposed approach. We start with the overview of MACON, and then describe the content modeling module and the user habit modeling module.

Model Overview

Given a post with both image and text in a photo sharing service, we aim to recommend the hashtags that the post owner is most likely to tag for the post. There are two important requirements for this task, i.e., (1) the hybrid modeling of both image and text, and (2) the modeling of users' tagging habits. In this work, we treat the problem as a multi-label classification problem, and handle the two requirements with a content modeling module and a user habit modeling module, respectively.

The overview of the proposed MACON model is depicted in Figure 1. As we can see, the input of the model consists of both text and image of the post, as well as the id of the user who posts the content. The text and image are transmitted to the content modeling module, and the user id serves as the input of the user habit modeling module. The content modeling module extracts coherent features (denoted as p) for image and text. The user habit modeling module indexes a few historical posts for the current user, and learns the tagging habit of this user based on these historical posts. This

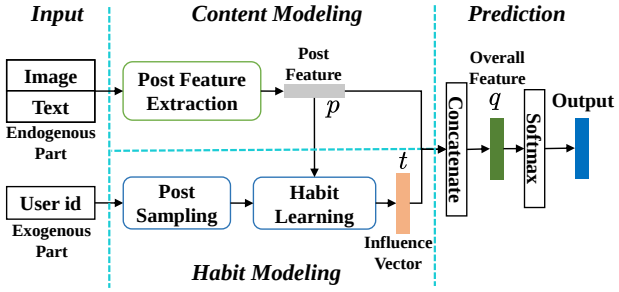


Figure 1: The overview of MACON.

module also takes the extracted content features p as input, so that the learned user habits can be properly applied to the current post. The output of the user habit modeling module is an influence vector (denoted as t). Finally, the post feature vector p from the content modeling module and the habit influence vector t from the user habit modeling module are concatenated to make the recommendations. More specifically, we employ a softmax layer to compute the probability of each hashtag, and a ranked list of top-K hashtags are returned for a given post.

Content Modeling

For content modeling, although there are many existing methods for separately modeling text and images, few efforts have been made on the interaction between text and images, especially in the context of photo sharing services. Here, we first separately extract features from text and image, and then model their interaction via parallel co-attention.

Modeling Text To generate text features, we embed each word with a vector x_i and the text can be represented as $[x_1, ..., x_N]$, where N is the maximum length of text. Next, we adopt LSTM to model the sequential nature of the text. That is, at each time step, the LSTM unit takes the word embedding x_i and the output of the previous unit h_{i-1} as input, and outputs h_i for the current word:

$$h_i = LSTM(x_i, h_{i-1}), \tag{1}$$

where $h_i \in \mathbb{R}^d, i = 1, 2, \cdots, N$, and d is the output dimension of LSTM. We omit the detailed equations for LSTM for brevity.

With LSTM networks, we can output a single feature vector for the input text. However, since we plan to apply co-attention to indicate the importance/weights of each word, we keep the feature vectors for all words. That is, denoting u as the text feature matrix, we have $u = [h_1, h_2, \cdots, h_N]$.

Modeling Images For image feature extraction, we utilize the pre-trained VGG-16 network (Simonyan and Zisserman 2014). Similar to the text feature extraction, we construct multiple feature vectors for an image by keeping the regional feature vectors. In particular, since the last pooling layer of VGG-16 is a $7 \times 7 \times 512$ tensor for 7×7 regions each of which is represented via a 512 dimensional vector, we

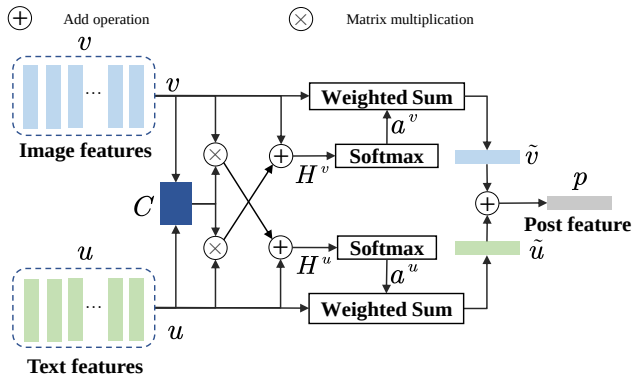


Figure 2: The content modeling module.

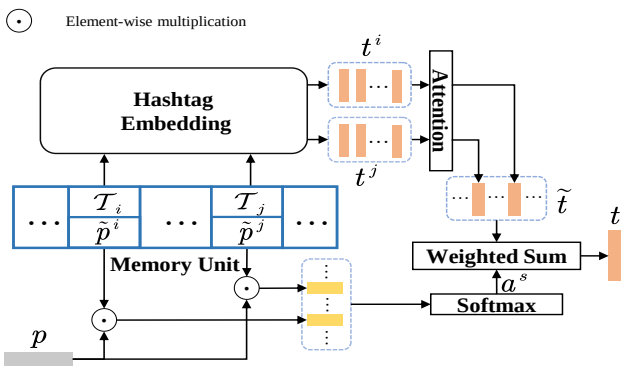


Figure 3: The user habit modeling module.

keep $M = 7 \times 7 = 49$ regional feature vectors for each image. Then, the feature matrix for an image can be written as $v^* = [v_1^*, v_2^*, \cdots, v_M^*]$ where $v_i^* \in \mathbb{R}^D, i = 1, 2, \cdots, M$ with $D = 512$.

For convenience, we further add a fully connected layer after the VGG network to convert each D-dimensional regional feature vector into a new vector that has the same dimension with the text feature vectors. As a result, the image feature matrix becomes $v = [v_1, v_2, \cdots, v_M]$ with $v_i \in \mathbb{R}^d, i = 1, 2, \cdots, M$.

Modeling Text–Image Interaction To extract the coherent features for a post, we employ a parallel co-attention mechanism (Lu et al. 2016) which can simultaneously learn the importance of each feature vector in both text features and image features towards the recommendation target. The overall architecture of the co-attention mechanism is shown in Figure 2, where we omit the details of the separate feature extractions as described in the previous two subsections.

Given $u \in \mathbb{R}^{d \times N}$ and $v \in \mathbb{R}^{d \times M}$, the parallel co-attention mechanism starts with defining an affinity matrix $C \in \mathbb{R}^{N \times M}$, whose element represents the similarity between the corresponding feature vector pair of u and v. Specifically, C is defined as

$$C = tanh(u^T W_b v), \tag{2}$$

where $W_b \in \mathbb{R}^{d \times d}$ denotes the correlation matrix to be learned.

Based on the affinity matrix, we next transfer the image and text features for each other to further exploit the correlations between image and text features. Take text features as an example, we can define the new text feature matrix $H^u \in \mathbb{R}^{d \times N}$ as follows,

$$H^u = tanh(W_u u + (W_v v)C^T), \tag{3}$$

where the image features v is multiplied by C^T and integrated into the text features, and $W_u, W_v \in \mathbb{R}^{d \times d}$ are parameters. By using this new feature matrix, the image features also serve as a guidance role for the attention learning of text. Similarly, we have the new image feature matrix $H^v \in \mathbb{R}^{d \times M}$,

$$H^v = tanh(W_v v + (W_u u)C). \tag{4}$$

Then, we calculate the attention weights of text and image as follows,

$$\begin{aligned} a^u &= softmax(W_{hu}^T H^u + b_{hu}), \\ a^v &= softmax(W_{hv}^T H^v + b_{hv}), \end{aligned} \tag{5}$$

where $W_{hu}, W_{hv} \in \mathbb{R}^d$ and $b_{hu}, b_{hv} \in \mathbb{R}$ are parameters.

Based on the attention weights, the global feature vectors for text and image could be represented as the weighted sum of the partial text and image feature vectors, i.e.,

$$\tilde{u} = \sum_{i=1}^{N} a_i^u u_i, \qquad \tilde{v} = \sum_{i=1}^{M} a_i^v v_i, \tag{6}$$

where a_i^u and a_i^v represent the attention weights corresponding to a certain word or an image region, respectively.

Finally, the post feature is represented as the sum of the two global feature vectors.

$$p = \tilde{u} + \tilde{v}. \tag{7}$$

Remarks. One of the existing tag recommendation efforts that consider both text and image is from CoA (Zhang et al. 2017). Our method differs from CoA in the following two aspects. First, we use parallel co-attention as image and text are equally informative for tagging in photo sharing services, while CoA chooses the alternative way for microblogs where the text is more informative. Second, we further model the user's tagging habit as we will show in the next subsection.

User Habit Modeling

For the user habit modeling module, we allocate a memory unit to store the habits for each user, and the memory unit can be indexed with user id. In particular, this module consists of two major steps. The first step samples a small number of users' historical posts and the corresponding hashtags as external memory. The second step learns the tagging habits in these historical posts and connects the habits with the current post to tag. The overall architecture of this module is shown in Figure 3.

Post Sampling Post sampling aims to select the historical posts where user habits can be learned from. There are various sampling strategies, and we list a few of them as follows. The first is *user-based random sampling*, which randomly samples the previous posts from the current user. Further considering the community effect, we can also adopt *community-based random sampling* where the posts from the friends of the current user can be sampled. When considering different strategies other than the random sampling, we can use *user-based temporal sampling* to select the more recent posts from the current user. For simplicity, we adopt the user-based random sampling in this paper, and leave the other sampling strategies as future work.

The sampled posts for each user are stored in the memory unit as shown in Figure 3 (blue array-like rectangles). Memory size is denoted as L. That is, L historical posts as well as their hashtags are sampled for each user. We constrain L to be a relatively small number as users may have posted a small number of posts.

Habit Learning For habit learning, we first extract the features of historical posts, and then measure the similarities between the current query post and the historical posts. Finally, we compute the influence vector t as a weighted sum of the hashtags in the memory unit where the weights are determined by the similarities.

For post feature extraction of historical posts, we use the approach as described in the content modeling module, and denote the feature vectors for these posts as $\tilde{p} = \{\tilde{p}^i | \tilde{p}^i \in \mathbb{R}^d, i = 1, 2, \cdots, L\}$. Notice that historical post features are not pre-trained or separately trained but trained together with the current post.

For the historical posts in the memory unit, we use $\mathcal{T}_i$ to denote the set of hashtags of the i-th historical post. Similar to the word embedding procedure in our text feature extraction, we introduce a hashtag embedding procedure to embed each set of hashtags $\mathcal{T}_i$ into $t^i \in \mathbb{R}^{d \times N_t}$, where N_t is the maximum size of hashtag sets and d is hashtag embedding dimension which is set equal to the word embedding dimension for convenience. We further add an attention mechanism to summarize each hashtag set into a single hashtag influence vector $\tilde{t^i}$. The equations are summarized as follows,

$$\begin{aligned} H^t &= tanh(W_t t^i), \\ a^t &= softmax(W_{ht}^T H^t + b_{ht}), \\ \tilde{t^i} &= \sum_{k=1}^{N_t} a_k^t t_k^i, \end{aligned} \tag{8}$$

where $W_t \in \mathbb{R}^{d \times d}, W_{ht} \in \mathbb{R}^d, b_{ht} \in \mathbb{R}$ are parameters, and t_k^i indicates the k-th column of matrix t^i. The hashtag influence vectors in the memory unit are denoted as $\tilde{t} = \{\tilde{t^i} | \tilde{t^i} \in \mathbb{R}^d, i = 1, 2, \cdots, L\}$.

Next, we measure the similarity between the query post and a historical post as follows,

$$s_i = tanh(p \odot \tilde{p}^i), \tag{9}$$

where $\odot$ means element-wise multiplication and s_i represents the correlation vector between the query post and the i-th historical post. Combining all correlation vectors, we have the similarity matrix $s = [s_1, s_2, \cdots, s_L]$. Based on s, we can compute the weights of each historical post as

$$a^s = softmax(W_s^T s + b_s), \tag{10}$$

where $W_s \in \mathbb{R}^d, b_s \in \mathbb{R}$ are parameters and $a^s \in \mathbb{R}^L$ is a vector containing the weights of historical posts.

Finally, the influence vector t can be computed as follows,

$$t = \sum_{i=1}^{L} a_i^s \tilde{t^i}. \tag{11}$$

This influence vector t reflects the user habits towards tagging the current post. It can be seen as a combined representation of the corresponding tags weighted by the similarities between the historical post and the current post.

Table 2: Statistics of the Instagram dataset. Ave_p is the average number of posts per user. Ave_h is the average number of hashtags per post.

#Posts	#Users	#Hashtags	Ave_p	Ave_h
624,520	7,497	3896	83.3	6.41

Training

We define the training objective function as below:

$$J = \frac{1}{|S|} \sum_{(p_i, \mathcal{T}_i) \in S} \sum_{z \in \mathcal{T}_i} -\log P(z|p_i), \tag{12}$$

where S is the training set, p_i and $\mathcal{T}_i$ are a post and its corresponding hashtag set, z is a hashtag in the hashtag set, and $P(z|p_i)$ is the softmax probability of choosing tag z for input post p_i.

Experimental Evaluations

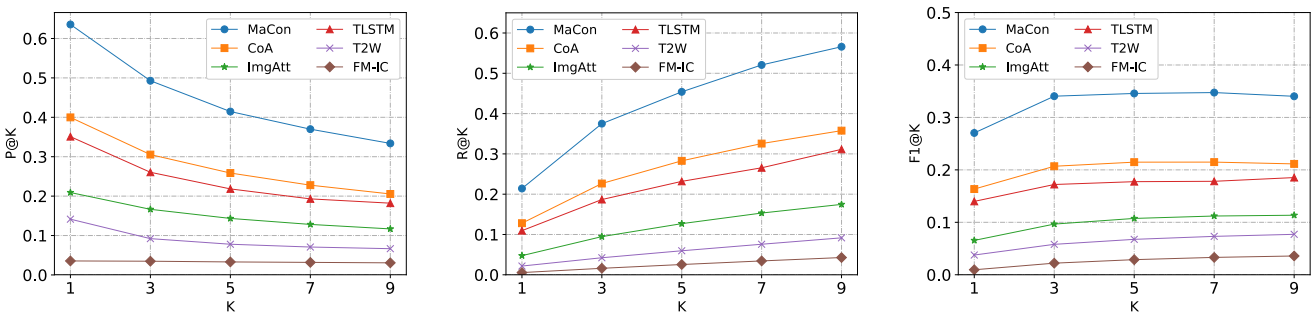
In this section, we present the experimental results.

Experimental Setup

Dataset We collect a dataset from Instagram. We first randomly select more than 15,000 users and crawl all their posts. Next, we remove some low frequency hashtags and words, and keep the posts that contain both image and text and at least one hashtag. The posts that have less than five words in the text are also removed. Finally, we remove the users as well as their posts if the user has less than 20 posts in the dataset. The ultimate dataset contains 624,520 posts from 7,497 users, and there are 3,896 unique hashtags and 212,000 distinct words. The statistics of the dataset are summarized in Table 2.

Compared Methods We compare the following methods:

- **Tag2Word (T2W)** (Wu et al. 2016): T2W is a supervised variant of topic modeling used for tag recommendation. It takes text as input.



(a) Precision (b) Recall (c) F1-Score

Figure 4: The effectiveness comparison results. The proposed MACON significantly outperforms the compared methods in the three evaluation metrics.

- **Topical attention based LSTM (TLSTM)** (Li et al. 2016): TLSTM also recommends hashtags for textual content, and it employs LSTM to extract textual features and integrates an attention mechanism by pre-trained topical distributions.
- **Image Attention (ImgAtt)** (Yang et al. 2016): ImgAtt is original proposed for visual question answering. It uses stacked attention networks to model both image and text. This model can be easily adapted for the hashtag recommendation problem.
- **Factorization Machine with Image Classification Feature (FM-IC)** (Nguyen, Wistuba, and Schmidt-Thieme 2017): FM-IC is a personalized hashtag recommendation method for images. It transfers the pre-trained image features into the factorization machine.
- **Co-Attention (CoA)** (Zhang et al. 2017): CoA is the state-of-the-art hashtag recommendation method for multimodal microblog posts with both text and image. This model applies the alternative co-attention mechanism to extract post features and then directly uses the features to make recommendations.
- **MACON**: MACON is the proposed method for hashtag recommendation. It adopts parallel co-attention to extract features from multimodal posts, and learns the user habits for better recommendations.

Evaluation Metrics The adopted evaluation metrics include precision (P), recall (R), and F1-score (F1). For example, $P@K$ represents the precision value when K hashtags are recommended for each post. For all three metrics, the higher the better.

Parameters and Reproducibility The default embedding dimension d is set to 300. For fairness, the embedding dimension for the other methods are also set as 300. For the user habit modeling module, we randomly select L posts for each user in order to model his/her habits. L is set to 2 unless otherwise stated. For images, we resize the images to 224×224 before feeding them into the pre-trained VGG-16 network. Our model is trained using stochastic gradient descent with the Adam optimizer. We also adopt dropout with dropout rate 0.75. For the dataset, we randomly select 90% data as the training set and use the rest 10% as the test set.[4]

Experimental Results

Effectiveness Comparisons We first compare the proposed method with the existing methods in Figure 4. The y-axis represents the precision, recall, and F1-score, respectively; the x-axis indicates the number of hashtags that the recommendation methods return.

We can observe from Figure 4 that, the proposed MACON significantly outperforms all the competitors on all the three evaluation metrics. Compared with the best competitor CoA, when K varies from 1 to 9, our method can achieve 12.8% - 23.6%, 8.6% - 20.8%, and 10.7% - 13.4% absolute improvements in terms of the precision, recall, and F1-score, respectively. The remarkable improvements demonstrate the effectiveness of our approach.

In the compared methods, CoA performs relatively better than the other competitors. One probable reason is that CoA models both text and image. Although ImgAtt also considers both image and text, it performs worse than CoA. This is probably due to the fact that ImgAtt is original designed for a different task, and it does not use co-attention. Both T2W and TLSTM are text-based methods, and TLSTM performs much better. This indicates the usefulness of modeling the text with LSTM networks. As to the personalized FM-IC method, it performs relatively poor, as it directly transfers the features from pre-trained models into the hashtag recommendation task.

Performance Gain Analysis Next, we analyze the performance gain of the proposed method. We consider three components of the proposed method, i.e., the text input, the image input, and the user habit modeling. To demonstrate the usefulness of user habit modeling module, we remove it and keep the content modeling module. We name the resulting method as MACON$_{t+i}$. To show the usefulness of the text input, we remove it and keep the image input and the user habit. The resulting method is denoted as MACON$_{i+h}$. Similarly, MACON$_{t+h}$ uses text input and user habit. The results

[4]The code of the proposed method is publicly available at https://github.com/SoftWiser-group/macon.

Table 3: The performance gain analysis. Both hybrid content modeling and user habit modeling are useful to improve recommendation accuracy.

top-K	MACON			MACON$_{t+i}$ (text+image)			MACON$_{t+h}$ (text+habit)			MACON$_{i+h}$ (image+habit)		
	P@K	R@K	F1@K	P@K	R@K	F1@K	P@K	R@K	F1@K	P@K	R@K	F1@K
1	0.636	0.214	0.270	0.384	0.060	0.105	0.596	0.196	0.248	0.560	0.171	0.221
3	0.493	0.375	0.341	0.300	0.142	0.193	0.457	0.340	0.311	0.441	0.305	0.289
5	0.415	0.454	0.346	0.259	0.204	0.228	0.388	0.418	0.320	0.379	0.382	0.305
7	0.370	0.521	0.347	0.231	0.255	0.243	0.360	0.503	0.337	0.337	0.437	0.307
9	0.334	0.566	0.340	0.210	0.298	0.247	0.311	0.517	0.314	0.307	0.481	0.305

of these variants are shown in Table 3, where we also report the MACON results for comparison.

We can first observe that MACON achieves better performance than MACON$_{t+i}$. This result indicates the importance of the habit modeling module. For example, when K varies from 1 to 9, the absolute improvements of MACON range from 9.3% - 16.5% in terms of F1-score. Moreover, the MACON$_{t+i}$ achieves up to 16.8% relative improvement compared with CoA. This result confirms our intuition that the adopted parallel co-attention is more suitable than the alternative way for photo sharing services where both image and text are equally informative for tagging.

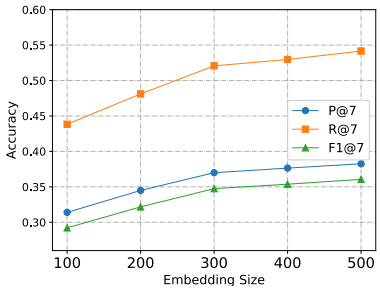
Second, MACON performs better than both MACON$_{t+h}$ and MACON$_{i+h}$ on all metrics. For example, when K varies from 1 to 9, the average relative F1-score improvements of MACON over MACON$_{t+h}$ and MACON$_{i+h}$ are 7.6% and 15.6%, respectively. This result indicates that both text and image are useful input for the hashtag recommendation problem in photo sharing services, and that our hybrid modeling of text and image can significantly improve the recommendation accuracy.

Third, comparing with the TLSTM method as shown in Figure 4, MACON$_{t+h}$ can also achieve 10.8% - 15.8% absolute improvements in terms of F1-score. Since MACON$_{t+h}$ can be seen as a combination of text modeling and user habit modeling, this result, again, shows the usefulness and the applicability of the proposed user habit modeling module.

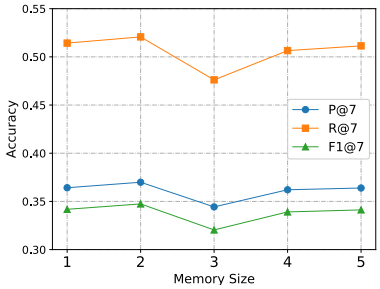
Parameter Sensitivity Study Finally, we study the effects of two parameters in our method, i.e., the embedding size d and the memory size L. The results are shown in Figure 5.

For the embedding size d, we vary it from 100 to 500. For simplicity, we only report the results when $K = 7$. Similar results are observed when K is set to other values. As we can see from Figure 5(a), MACON with larger embedding size tends to have better recommendation accuracy. For example, increasing d from 100 to 300 has witnessed a notable performance improvement. However, when d grows from 400 to 500, only slight performance gain is observed. Considering that increasing embedding size would cost more time and memory, we fix the embedding size as $d = 300$ in this work.

For the memory size L, which means how many historical posts are selected to model the current user's tagging habit, we vary it from 1 to 5. We still report the results when $K = 7$ in Figure 5(b). As we can see from the figure, the performance remains relatively stable as L varies. In our experiments, we fix the memory size as $L = 2$.



(a) The effect of the embedding size d



(b) The effect of the memory size L

Figure 5: The parameter sensitivity study. The performance of MACON increases as d increases, and stays relatively stable as L varies. We fix $d = 300$ and $L = 2$ in this work.

Conclusions

In this paper, we propose a novel approach for hashtag recommendation in photo sharing services. The established approach consists of two key components: post content modeling and user habit modeling. For the former, we employ a parallel co-attention neural network to coherently learn the features of both image and text. For the latter, we introduce an external memory unit to store historical posts and learn the influence of the current user's tagging habit. Experimental evaluations are conducted on the crawled Instagram dataset. The results demonstrate that the proposed method can achieve significantly better performance than the existing methods, and that the two components of post content modeling and user habit modeling play important roles to improve the recommendation accuracy. In the future, we plan to further explore the user habit modeling module with different sampling strategies by, for example, incorporating the community effect and the temporal effect.

Acknowledgments. This work is supported by the National Key Research and Development Program of China (No. 2017YFB1001801), the National Natural Science Foundation of China (No. 61690204, 61672274, 61702252), the Huawei Innovation Research Program (No. HO2018085291), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. Hanghang Tong is partially supported by NSF (IIS-1651203, IIS-1715385 and IIS-1743040), and DHS (2017-ST-061-QA0001). Xiaohui Yan is partially supported by the National Natural Science Foundation of China (No. 61502447).

References

Fang, X.; Pan, R.; Cao, G.; He, X.; and Dai, W. 2015. Personalized tag recommendation through nonlinear tensor factorization using gaussian kernel. In *AAAI*, 439–445.

Feng, W., and Wang, J. 2012. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *SIGKDD*, 1276–1284. ACM.

Gong, Y., and Zhang, Q. 2016. Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*, 2782–2788.

Gong, Y.; Jia, Y.; Leung, T.; Toshev, A.; and Ioffe, S. 2013. Deep convolutional ranking for multilabel image annotation. *arXiv*.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV* 106(2):210–233.

Guan, Z.; Bu, J.; Mei, Q.; Chen, C.; and Wang, C. 2009. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *SIGIR*, 540–547. ACM.

Hasan, M.; Agu, E.; and Rundensteiner, E. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD Workshop on Health Informatics*.

Highfield, T., and Leaver, T. 2015. A methodology for mapping instagram hashtags. *First Monday* 20(1):1–11.

Huang, H.; Zhang, Q.; Gong, Y.; and Huang, X. 2016. Hashtag recommendation using end-to-end memory networks with hierarchical attention. In *COLING*, 943–952.

Hwang, S. J., and Grauman, K. 2012. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International journal of computer vision* 100(2):134–153.

Krestel, R.; Fankhauser, P.; and Nejdl, W. 2009. Latent dirichlet allocation for tag recommendation. In *RecSys*, 61–68.

Li, Y.; Liu, T.; Jiang, J.; and Zhang, L. 2016. Hashtag recommendation with topical attention-based lstm. In *COLING*, 3019–3029.

Lim, K. W., and Buntine, W. 2014. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *CIKM*, 1319–1328.

Liu, D.; Hua, X.-S.; Yang, L.; Wang, M.; and Zhang, H.-J. 2009. Tag ranking. In *WWW*, 351–360. ACM.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 289–297.

Nguyen, H. T.; Wistuba, M.; and Schmidt-Thieme, L. 2017. Personalized tag recommendation for images using deep transfer learning. In *ECML/PKDD*, 705–720.

Qian, X.; Liu, X.; Zheng, C.; Du, Y.; and Hou, X. 2013. Tagging photos using users' vocabularies. *Neurocomputing* 111:144–153.

Rawat, Y. S., and Kankanhalli, M. S. 2016. Contagnet: Exploiting user context for image tag recommendation. In *MM*, 1102–1106. ACM.

Rendle, S., and Schmidt-Thieme, L. 2010. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM*, 81–90. ACM.

Sedhai, S., and Sun, A. 2014. Hashtag recommendation for hyperlinked tweets. In *SIGIR*, 831–834.

Sigurbjörnsson, B., and Van Zwol, R. 2008. Flickr tag recommendation based on collective knowledge. In *WWW*, 327–336.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv*.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *NIPS*, 2440–2448.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2285–2294.

Wang, H.; Shi, X.; and Yeung, D.-Y. 2015. Relational stacked denoising autoencoder for tag recommendation. In *AAAI*, 3052–3058.

Wei, Y.; Xia, W.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; and Yan, S. 2014. Cnn: single-label to multi-label. *arXiv*.

Weston, J.; Chopra, S.; and Adams, K. 2014. # tagspace: Semantic embeddings from hashtags. In *EMNLP*, 1822–1827.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *CoRR* abs/1410.3916.

Wu, Y.; Yao, Y.; Xu, F.; Tong, H.; and Lu, J. 2016. Tag2word: Using tags to generate words for content based tag recommendation. In *CIKM*, 2287–2292.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *CVPR*, 21–29.

Zhang, Q.; Wang, J.; Huang, H.; Huang, X.; and Gong, Y. 2017. Hashtag recommendation for multimodal microblog using co-attention network. In *IJCAI*, 3420–3426.

Zhu, S.; Aloufi, S.; and El Saddik, A. 2015. Utilizing image social clues for automated image tagging. In *ICME*, 1–6.