

Hashtag Recommendation for Multimodal Microblog Using Co-Attention Network

Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, Yeyun Gong

School of Computer Science, Fudan University

Shanghai Key Laboratory of Intelligent Information Processing

825 Zhangheng Road, Shanghai, P.R. China

{qz, wangjiawen16, huanghr15, xjhuang, yygong12}@fudan.edu.cn

Abstract

In microblogging services, users usually use hashtags to mark keywords or topics. Along with the fast growing of social network, the task of automatically recommending hashtags has received considerable attention in recent years. Previous works focused only on the use of textual information. However, many microblog posts contain not only texts but also the corresponding images. These images can provide additional information that is not included in the text, which could be helpful to improve the accuracy of hashtag recommendation. Motivated by the successful use of the attention mechanism, we propose a co-attention network incorporating textual and visual information to recommend hashtags for multimodal tweets. Experimental results on the data collected from Twitter demonstrated that the proposed method can achieve better performance than state-of-the-art methods using textual information only.

1 Introduction

In recent years, microblogging, like Twitter and Sina Weibo, has become one of the most popular services for information generation and diffusion, as well as social interaction among the various social media outlets. According to the quarterly report released by Twitter¹, there are 317 million monthly active users. The users write texts with limited number of characters to record life or express their emotion. Therefore, microblogs have been widely used as sources for public opinion analysis [Bermingham and Smeaton, 2010], prediction [Bollen *et al.*, 2011], and many other applications [Sakaki *et al.*, 2010; Becker *et al.*, 2010; Guy *et al.*, 2013]. Microblogs contain a form of metadata tag (hashtag), which is a string of characters prefixed with the symbol (#). Hashtags are used to mark keywords or topics within a microblog and have proven to be useful for many applications, including microblog retrieval [Efron, 2010], query expansion [Bandyopadhyay *et al.*, 2012], and sentiment analysis [Wang *et al.*, 2011].

¹Twitter Q3 report 2016: <https://investor.twitterinc.com/results.cfm>



Figure 1: An example of multimodal tweets. Without visual information, we can hardly predict the correct tag: #dog.

However, despite their proven success, relatively few microblogs include hashtags labeled by their users. Hence, the task of automatically recommending hashtags has become an important research topic and has received considerable attention in recent years. Various models have been adopted for this task using kinds of features [Ohkura *et al.*, 2006; Heymann *et al.*, 2008], collaborative filtering [Kywe *et al.*, 2012], generative models [Krestel *et al.*, 2009; Ding *et al.*, 2013] and deep neural networks [Gong and Zhang, 2016]. Although some research has been done on this topic, most of the studies have focused only on the use of textual information. However, according to the data retrieved from Twitter, we observe that more than 30% of tweets contain not only text but also images. Hence, it is not easy to correctly recommend hashtags through approaches designed to use only textual information. Figure 1 illustrates a multimodal microblog with the hashtag #dog. There is no information about dog in this tweet. With only textual information, we may extract the hashtag #birthday. However, the hashtag #dog is hardly to be identified.

To address this issue, we present a multimodal model to combine textual and visual information together. Some

previous works simply combine the image feature vector and the text feature vector [Antol *et al.*, 2015]. However, the correct hashtags are often only related to a small part of images or texts. Hence, using a global vector to represent the image or text may lead to suboptimal results due to the noises made by the irrelevant part of images or texts. Motivated by the work [Yang *et al.*, 2015; Ma *et al.*, 2015] on image QA task and [Vinyals *et al.*, 2015] on image captioning, we introduce an attention mechanism to conduct the task of hashtag recommendation. Attention mechanism allows the model to focus on specific parts of the input. Since image captioning and image QA are mainly focused on image processing, these works mentioned above only intend to model image features by taking image attention into consideration. However, textual information is an essential part of hashtag recommendation tasks for multimodal microblogs. In addition to the textual information, the image also can guide feature extraction of the text. In this work, we introduce a co-attention network that takes the mutual influence of both text and image into consideration.

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

- We introduce an integrated framework of visual and textual information for hashtag recommendation tasks.
- We propose a co-attention network that incorporates tweet-guided visual attention and image-guided textual attention.
- Experimental results using a dataset collected from Twitter demonstrate that the proposed method can achieve significantly better performance than current state-of-the-art methods.

2 Related Works

2.1 Hashtag Recommendation

Because of the usefulness of hashtag recommendation, many methods have been proposed from different perspectives. [Krestel *et al.*, 2009] introduced Latent Dirichlet Allocation to elicit a shared topical structure from the collaborative tagging efforts of multiple users for recommending hashtags. Based on the observation that similar webpages tend to have the same tags, [Lu *et al.*, 2009] proposed a method taking both tag information and page content into account to achieve the task. [Ding *et al.*, 2013] proposed the use of a translation process to model this task. They extended this translation-based method and introduced a topic-specific translation model to process various meanings of words in different topics. In [Tariq *et al.*, 2013], discriminative-term-weights were used to establish topic-term relationships, of which users perception were learned to suggest suitable hashtags for users. [Gong and Zhang, 2016] adopted CNNs with an attention mechanism to perform this task. They added an extra channel to take trigger words into consideration.

The task of hashtag suggestion/recommendation for images is also related to this work and has been studied from various aspects. Most of these works give much attention to the tags annotated by users through social media services such as Flickr, Zoomr, and so on. The tags annotated in these services are like labels that are added to a photo to make it easier to find later. Previous works mainly focused on recommending tags that are good descriptors of the photo itself, whereas hashtags are usually referred to more abstract concepts. [Sigurbjörnsson and Van Zwol, 2008] studied the tag recommendation task for images. Their approach was based on the statistics of Flickr annotation patterns and tag co-occurrence statics. When a user submits a photo and enters some tags, an ordered list of candidate tags is derived for each of those entered tags. Hence, it cannot be directly applied to images only. In [Garg and Weber, 2008], the problem of personalized, interactive tag recommendation was also studied based on the statistics of the tags co-occurrence. In [Li and Snoek, 2013], ensembles of Support Vector Machines per tag was used to classify tag relevance.

From the brief descriptions given above, we can observe that most of the previous works focused on either textual information or visual features. In this work, the proposed method incorporates both textual and visual information.

2.2 Multimodal Model

There is a large quantity of literature on multimodal models. Early works usually focused on modeling the relation between an image and text. Recently, the association between an image and text has been studied for automatic image captioning and image QA tasks. [Chen and Zitnick, 2014] used a recurrent visual memory to aid in both sentence generation and visual feature reconstruction. [Vinyals *et al.*, 2015] first extracted high-level image features and then fed them into an LSTM to generate captions. [Antol *et al.*, 2015] used an LSTM to encode a question and then combined the question with an image using element-wise multiplication. [Yang *et al.*, 2015] used an attention mechanism to query an image multiple times to infer the answer progressively. However, as previously mentioned in the Introduction section, these works did not consider the guiding significance of an image on textual feature extraction.

Attention mechanism allows models to focus on specific parts of a visual or textual input and has been successfully used in various multimodal models. In this work, we adopt the mechanism to select important information from the input tweets and images.

3 The Proposed Models

Given a tweet with corresponding images, our task is to automatically generate proper hashtags for the tweet. To conduct this task on multimodal tweets, we formulate the task as a multi-class classification problem. The overall architecture of the model is shown in Figure 2. The input to our network is an image and a tweet comprised of a variable-length sequence of words. The output is a vector, and each dimension of the vector represents the probability of a hashtag. To ease understanding, we describe our model in

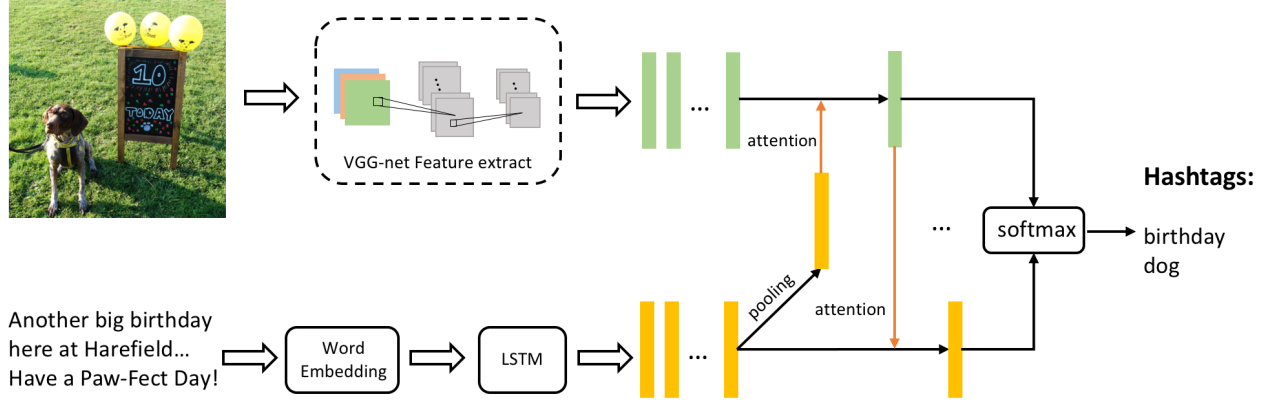


Figure 2: The graphical representation of the proposed model.

three parts. The feature extraction is described in Sec 3.1. The co-attention network and hashtag prediction are described in Sec 3.2 and Sec 3.3, respectively.

3.1 Feature Extraction

Image feature extraction

We use a pretrained 16-layer VGGNet [Simonyan and Zisserman, 2014] to extract an image feature map. We first rescale images to 224×224 . Unlike previous works that use a global vector as the image feature, we want spatial features of different regions which contain more information of the original image. We divide an image into an $N \times N$ grid, and then use VGGNet to extract a D -dimensional feature vector for each region of grids. Therefore, an image could be represented as $v_I = \{v_i | v_i \in \mathbb{R}^D, i = 1, 2, \dots, m\}$, where $m = N \times N$ is the number of regions in the image, which is equal to 49 in our case and v_i is a 512-dimensional feature vector corresponding to each region i . For convenience of calculation, we use a single layer perceptron to convert each image vector into a new vector that has the same dimension as the tweet feature vector.

Text feature extraction

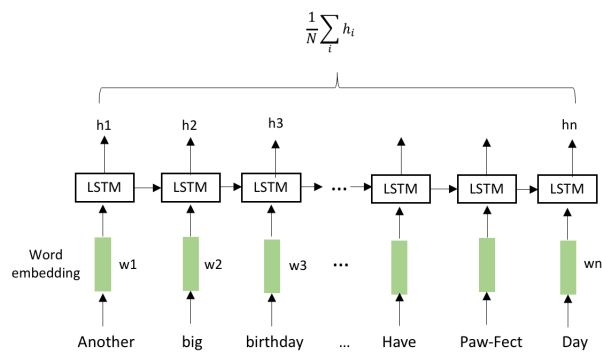


Figure 3: Text feature extraction using LSTM based model.

Each word of a given tweet t is first represented as a one-hot vector in the size of the vocabulary. Then, each one-hot vector is embedded into a real-valued word vector x_i distributed in a continuous space. We sum up the embedded vectors to obtain a sentence-level tweet representation: $t = x_1, x_2, \dots, x_T$, where T is the maximum length of the tweets. In our work, sentences with length less than T are padded with zeros.

Because the LSTM has shown good performance in understanding text and has been widely used in recent years, we employ it to generate text feature. The process of text feature extraction is shown in Figure 3. At each time step, the LSTM unit takes an input vector (word embedding vector in our case) x_t and outputs a hidden state h_t , using input gate i_t , memory cell c_t , forget gate f_t , and output gate o_t . The details are illustrated as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \quad (1)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \quad (2)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \quad (3)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \quad (4)$$

$$h_t = o_t \tanh(c_t), \quad (5)$$

We construct the tweet feature matrix v_T by combining the h_t states at every time step: $v_T = \{h_i | h_i \in \mathbb{R}^d, i = 1, 2, \dots, T\}$, where d is the LSTM's dimension and T is the maximum length of a tweet. In particular, the semantic information h_{ave} used in Sec 3.2 is obtained by averaging the LSTM states h_t over all of the time steps.

3.2 Co-Attention Network

Based on the notations provided in Sec 3.1, we have the image feature matrix v_I and the text feature matrix v_T . Because tweets and images are not equally important in the hashtag recommendation task, we propose a co-attention network that generates tweet attention and image attention sequentially. Because text is the main part of the hashtag recommendation task for tweets, we first attend to images based on tweets.

Tweet-guided visual attention

In most cases, a hashtag is only related to a specific region of the input image. For example, in Figure 1, there are many objects in the image: balloons, a dog, a blackboard, and grass, but the hashtag #dog is only related to the little dog in the image. Hence, instead of using a global vector as an image feature, we divide an image into grids and extract the feature vector of each grid to obtain a feature matrix v_I . We then use a tweet-guided attention layer to filter out noises and find regions that are relevant to the corresponding hashtags.

As mentioned in Sec 3.1, we first use an average pooling operation to summarize the tweet features into a single vector h_{ave} . Next, we propose an image attention based on the tweet summary h_{ave} . We use a single-layer neural network to combine the tweet and image features and then use a softmax layer to generate an image attention distribution. The equation for these operations is as follows:

$$\begin{aligned} h_I &= \tanh(W_{v_I} v_I \odot W_{v_T} v_T), \\ p_I &= \text{softmax}(W_{p_I} h_I + b_{p_I}), \end{aligned} \quad (6)$$

where $v_I \in \mathbb{R}^{d \times m}$ and $v_T \in \mathbb{R}^d$, m is the number of regions in an image, and d is the dimension of the representation. W_{v_I} , W_{v_T} , W_{p_I} are parameters, and $W_{v_I}, W_{v_T} \in \mathbb{R}^{k \times d}$, $W_{p_I} \in \mathbb{R}^{1 \times k}$. Therefore, $p_I \in \mathbb{R}^m$, which corresponds to the attention probability of each region, is an m -dimensional vector. In addition, we use \odot to denote the combination of the image feature matrix and tweet feature vector, which is obtained by concatenating each column of the matrix by the vector.

Based on the attention probability p_i of each image region i , the new representation of the image is constructed as the weighted sum of the image vector,

$$\tilde{v}_I = \sum_i p_i v_i. \quad (8)$$

We then use the new image representation \tilde{v}_I to guide the textual attention.

Image-guided textual attention

Compared to models that only use text to guide image attention, the co-attention network model constructs a more informative representation by considering the mutual influence of both text and image. The process of image guided textual attention is similar to visual attention. In order to obtain the textual attention distribution, we use the new representation \tilde{v}_I of image to query the original text feature matrix v_T . We then generate a new representation \tilde{v}_T for text based on the probability distributions. The detail is illustrated as follows:

$$h_T = \tanh(W_{\tilde{v}_I} \tilde{v}_I \odot W_T v_T), \quad (9)$$

$$p_T = \text{softmax}(W_{p_T} h_T + b_{p_T}), \quad (10)$$

$$\tilde{v}_T = \sum_i p_i v_i, \quad (11)$$

where $v_T \in \mathbb{R}^{d \times T}$ and $\tilde{v}_I \in \mathbb{R}^d$, T is the maximum length of tweets, and d is the dimension of the representation. $W_{\tilde{v}_I}$,

W_T , W_{p_T} , b_{p_T} are parameters, and $W_{\tilde{v}_I}$, $W_T \in \mathbb{R}^{k \times d}$, $W_{p_T} \in \mathbb{R}^{1 \times k}$. Therefore, $p_T \in \mathbb{R}^T$, which corresponds to the attention probability of each word of a tweet.

Stacked co-attention network

For more complex tweets, we can try to explore some subtle relationships among text and images by iteratively querying the original feature matrixes of images and texts using the newly generated representations. Formally, the formula can be summarized as follows: for the k -th (where k is greater than or equal to 2) co-attention layer, we respectively compute the distribution of visual and textual attention and generate a new representation for the input image and text based on the attention probability. The new query vector is formed by adding the new feature vector to the previous vector, the equation for image attention is as follows:

$$q_I^k = \tilde{v}_I^k + q_I^{k-1}, \quad (12)$$

where q_I^1 is initialized to be \tilde{v}_I . The query vector q_T^k for textual attention is as same.

3.3 Prediction

Finally, the hashtags are predicted using a multi-class classification. We adopted a single-layer softmax classifier with cross-entropy loss, and the input is a combination of the features generated from both attention operations. The final vector $f = \tilde{v}_T + \tilde{v}_I$, and we achieve the scores of the hashtags for the d th tweet using:

$$P(y^d = a | h^d, \theta) = \frac{\exp(\theta^{(a)T}(W_f f + b_f))}{\sum_{j \in A} \exp(\theta^{(j)T}(W_f f + b_f))}, \quad (13)$$

where W_f, b_f, θ are parameters, A is the set of candidate hashtags.

According to the scores output from the last softmax layer, we can get a ranked list of hashtags for each tweet and recommend the top-ranked hashtags to users.

3.4 Training

In our work, the training objective function is:

$$J = \sum_{(t_p, \theta, h_p) \in S} -\log p(h_p | t_p; \theta), \quad (14)$$

where h_p is the hashtag for tweet t_p , and S is the training set.

The parameters in our model is:

$$\theta = \{\mathbf{W}, \mathbf{M}^i, \mathbf{M}^t, \mathbf{M}_{att}^i, \mathbf{M}_{att}^t\}, \quad (15)$$

where \mathbf{W} are words embeddings; \mathbf{M}^i and \mathbf{M}^t are the parameters of the feature extraction of images and texts respectively; \mathbf{M}_{att}^i and \mathbf{M}_{att}^t are the parameters of textual and visual attention layers; the rest parameters belong to the fully connected layer.

The parameters were trained using stochastic gradient descent with the adam [Kingma and Ba, 2014] update rule. Dropout regularization [Srivastava *et al.*, 2014] has proved to be an effective method and is used in our work.

4 Experiments

4.1 Dataset and Setup

We started by using Twitters API² to collect public tweets from randomly selected users. The collection contained 282.2 million microblogs published by 1.1 million users. From these microblogs, we extracted those that contained both images and hashtags. In this step, 2.27 million microblogs were extracted. Since some hashtags rarely occur, we filtered out the hashtags whose frequencies were very low in our corpus. Finally, the collection we constructed contained 402,782 tweets with corresponding images and hashtags of high frequency. The detailed statistics are shown in Table 1. The unique number of hashtags in the corpus was 3,292, and the average number of hashtags per tweet was 1.17. We split the dataset into a training set and a test set, with a ratio of 8:2, and randomly selected 10% of the training set as the development set.

#Tweets	#Images	#Hashtags	Ave_h
402,782	402,782	3292	1.17

Table 1: Statistics of the evaluation dataset. Ave_h represents the average number of manually labeled hashtags per tweet.

For text words, we filtered out the stop words and low-frequency words. The constructed word vocabulary contained 278,000 distinct words. For images, we downloaded images from the retrieved urls and rescaled them to 224×224 . Then we fed them into a pre-trained VGG-16 network. The outputs of the last pooling layer of VGGnet were extracted as the image feature.

We used precision (P), recall (R), and the F1-score (F1) to evaluate the performance. The number of recommended hashtags for each tweet is denoted as k , where $k = \{1, 2, 3, 4, 5\}$, and the P, R, and F1 at the k result are denoted as P_k , R_k , and $F1_k$, respectively.

4.2 Baseline

For comparison with the proposed model, we evaluated the following methods on the constructed corpus:

- **Naive Bayes (NB):** The hashtag recommendation task is formalized as a classification task. We applied NB to model the posterior probability of each hashtag given the textual and visual information of the microblogs.
- **Support Vector Machine (SVM):** We also followed the method proposed in [Chen *et al.*, 2008], which used an SVM to solve the tag recommendation problem.
- **TTM:** TTM was proposed by [Ding *et al.*, 2013] for hashtag recommendation using only textual information. The topical translation model was used to recommend hashtags.
- **CNN-Attention:** CNN-Attention was proposed by [Gong and Zhang, 2016] and is a convolutional neural network architecture that uses the attention

²<https://dev.twitter.com/>

Methods	Precision	Recall	F1
NB	0.090	0.081	0.085
SVM	0.169	0.155	0.161
TTM	0.195	0.195	0.195
CNN-Attention	0.237	0.236	0.237
Image-Att	0.266	0.241	0.253
Co-Att-2layer	0.292	0.268	0.279
Co-Attention	0.311	0.286	0.298

Table 2: Results of different methods on the test data set.

mechanism to incorporate trigger words, and it was the state-of-the-art method for this task.

- **Image-Att:** Image-Att is a variant of our proposed model, which only uses text information to generate visual attention distribution. Similar work was done by [Yang *et al.*, 2015].
- **Co-Att-2layer:** This is also a variant of our proposed model. We applied a stacked two-layer co-attention network to model the images and tweets.

4.3 Results and Discussion

We evaluated the proposed method from the following perspectives: 1) comparing the proposed method to state-of-the-art methods using a real world dataset and 2) identifying the impacts of the parameters.

Table 2 shows a comparison of the proposed method to the state-of-the-art discriminative and generative methods on the constructed evaluation collection. The three metric results listed were obtained when we recommended the top one hashtag for each tweet. Based on the results, we observed that the proposed method is better than the other methods. Discriminative methods achieved worse results than generative methods. The results of the proposed methods with and without visual information demonstrated that visual information can benefit both P and R.

Observing the comparisons of the "CNN-attention" and our proposed model, it is clear that images can significantly improve the performance of hashtag recommendation task. Since there are more and more users prefer to publish a tweet with corresponding images, finding an effective way to incorporate both textual and visual information posted is very meaningful.

For the lower performances, we believe there are several factors that led to this result, including the large number of labels. The dataset itself may be very confusing and difficult to distinguish, since all methods resulted in a low score in the evaluation dataset. In addition, the diversity of the image types and sources may impact methods using visual information. The VGG network we used for image feature extraction was pre-trained on an ILSVRC dataset, which includes images of 1,000 classes, most of which were outdoor scenes. However, the images crawled from Twitter were more irregular and some of them were graffiti, selfies, or even just screenshots of smart phones. Therefore, simply adopting a pre-trained weight may have caused deviations.

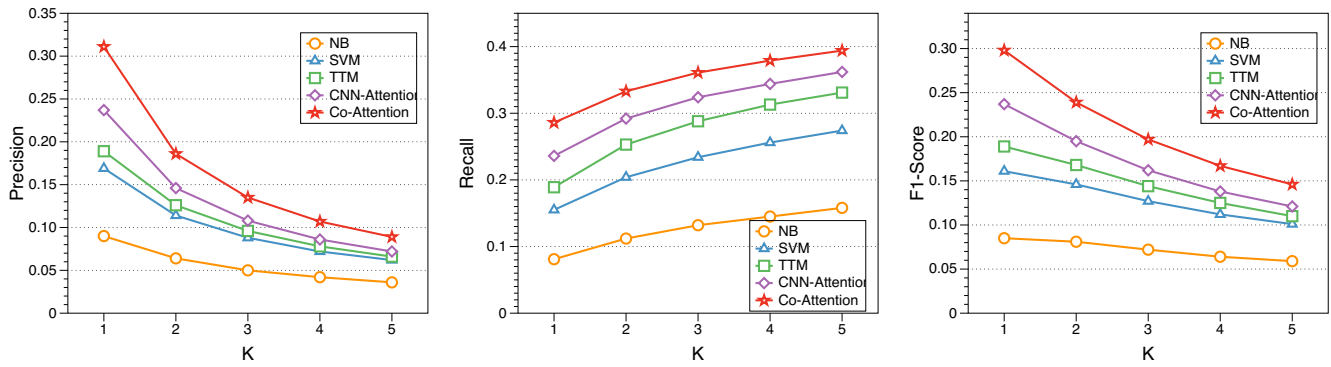


Figure 4: Precision, Recall, and F1-Score with different number of recommendation hashtags.

Methods	dim	Precision	Recall	F1
Image-Att	100	0.216	0.196	0.205
	300	0.245	0.223	0.233
	500	0.266	0.241	0.253
Co-Attention	100	0.267	0.245	0.256
	300	0.304	0.280	0.291
	500	0.311	0.286	0.298

Table 3: Results of variant methods of our proposed model and parameter Influence on the evaluation collection.

Figure 4 shows the P, R, and F1 of models with different numbers of recommended hashtags on the evaluation dataset. Each point of the curve represents the number of hashtags recommended ranging from 1 to 5. Obviously, P decreases and R increases as the number of recommended hashtags increases, and we obtained the highest F1 when recommending the top one hashtag. The curve that is the highest on the graph indicates the best performance. From the results, we can see that the performance of our proposed method is the highest of all the methods. This also indicates that the proposed method was significantly better than the state-of-the-art methods.

We also compare the precision, recall and F1-score of models with variants of our proposed method. Image-att only use text information to generate visual attention distribution. The results listed in Table 3 show that co-attention model achieve better results for the model that only use tweet-guided visual attention. The result of the model using a stacked two-layer co-attention work is slightly lower than the one layer co-attention network. We believe the short length of tweets limit the extent of layers. Since increasing number of layers of the network will make the model more complex and require more training time, and the result is not significant improved, we do not try more co-attention layers model.

4.4 Parameter Influence

The result listed in Table 3 also shows the contribution of embedding dimension to the performance. Higher embedding dimension leads to a higher performance. We can see that when the dimension equals to 100, the result is very poor. Our proposed model performed well with a high embedding

dimension. The performance improved significantly when the dimension changed from 100 to 300 and improve slightly from 300 to 500. The size of the embedding dimension represents the expression ability of each word, and a higher dimension can enhance the text feature expression ability.

5 Conclusion

In this paper, we propose and study a novel task, recommending hashtags for multimodal microblogs. We convert the hashtag suggestion task into a multi-way classification and introduce a co-attention network for this task. The proposed model combines textual and visual information. And we adopt attention mechanism to obtain more informative representation for both text and image. Since tweets and images are not equally important in hashtag recommendation task, we propose the co-attention network which generates textual attention and visual attention sequentially. We also constructed a large data collection retrieved from live microblog services to evaluate the effectiveness of our model. Experimental results showed that the proposed method is capable to achieve better performance than state-of-the-art methods using textual information only.

Acknowledgments

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

References

- [Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiaseen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.
- [Bandyopadhyay *et al.*, 2012] Ayan Bandyopadhyay, Kripabandhu Ghosh, Prasenjit Majumder, and Mandar Mitra. Query expansion for microblog retrieval. volume 1, pages 368–380. Inderscience Publishers Ltd, 2012.

- [Becker *et al.*, 2010] Hila Becker, Mor Naaman, and Luis Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of WSDM '10*, 2010.
- [Birmingham and Smeaton, 2010] Adam Birmingham and Alan F. Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of CIKM '10*, 2010.
- [Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
- [Chen and Zitnick, 2014] Xinlei Chen and C Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014.
- [Chen *et al.*, 2008] Hong-Ming Chen, Ming-Hsiu Chang, Ping-Chieh Chang, Ming-Chun Tien, Winston H Hsu, and Ja-Ling Wu. Sheepdog: group and tag recommendation for flickr photos by automatic search-based learning. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 737–740. ACM, 2008.
- [Ding *et al.*, 2013] Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. Learning topical translation model for microblog hashtag suggestion. In *IJCAI. Citeseer*, 2013.
- [Efron, 2010] Miles Efron. Hashtag retrieval in a microblogging environment. In *Proceedings of SIGIR '10*, 2010.
- [Garg and Weber, 2008] Nikhil Garg and Ingmar Weber. Personalized, interactive tag recommendation for flickr. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 67–74. ACM, 2008.
- [Gong and Zhang, 2016] Yuyun Gong and Qi Zhang. Hash-tag recommendation using attention-based convolutional neural network. In *IJCAI 2016, Proceedings of the 26rd International Joint Conference on Artificial Intelligence, New York City, USA, 2016*.
- [Guy *et al.*, 2013] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *Proceedings of WWW '13*, 2013.
- [Heymann *et al.*, 2008] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538. ACM, 2008.
- [Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [Krestel *et al.*, 2009] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 61–68. ACM, 2009.
- [Kywe *et al.*, 2012] Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu. On recommending hashtags in twitter networks. In *International Conference on Social Informatics*, pages 337–350. Springer, 2012.
- [Li and Snoek, 2013] Xirong Li and Cees GM Snoek. Classifying tag relevance with relevant positive and negative examples. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 485–488. ACM, 2013.
- [Lu *et al.*, 2009] Yu-Ta Lu, Shou-I Yu, Tsung-Chieh Chang, and Jane Yung-jen Hsu. A content-based method to enhance tag recommendation. In *IJCAI*, volume 9, pages 2064–2069, 2009.
- [Ma *et al.*, 2015] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. *arXiv preprint arXiv:1506.00333*, 2015.
- [Ohkura *et al.*, 2006] Tsutomu Ohkura, Yoji Kiyota, and Hiroshi Nakagawa. Browsing system for weblog articles based on automated folksonomy. In *Proceedings of the WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, at WWW*, volume 2006, 2006.
- [Sakaki *et al.*, 2010] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW '10*, 2010.
- [Sigurbjörnsson and Van Zwol, 2008] Börkur Sigurbjörnsson and Roelof Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336. ACM, 2008.
- [Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [Tariq *et al.*, 2013] Amara Tariq, Asim Karim, Fernando Gomez, and Hassan Foroosh. Exploiting topical perceptions over multi-lingual text for hashtag suggestion on twitter. In *FLAIRS Conference*, 2013.
- [Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.
- [Wang *et al.*, 2011] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of CIKM '11*, 2011.
- [Yang *et al.*, 2015] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015.