USER CONDITIONAL HASHTAG RECOMMENDATION FOR MICRO-VIDEOS

Shang Liu, Jiayi Xie, Cong Zou, Zhenzhong Chen[*]

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
{shangliu, xjyxie, congzou, zzchen}@whu.edu.cn

ABSTRACT

When a user tend to publish a micro-video, hashtag recommendation aims to suggest hashtags that can reflect the theme or contents of the micro-video, and meet the user tagging preference as well. In this paper, we show how user profile and historical hashtags combined with micro-video representations can be used to perform hashtag recommendation. Specifically, a User-guided Hierarchical Multi-head Attention Network (UHMAN) is proposed to attend both image-level and video-level representations of micro-videos with user side information. We evaluate the proposed model on the dataset collected from micro-video sharing platform Musical.ly. The experimental results demonstrate the effectiveness of the proposed method.

***Index Terms*— ** Attention network, Hashtag recommendation, User modeling, Micro-video representation

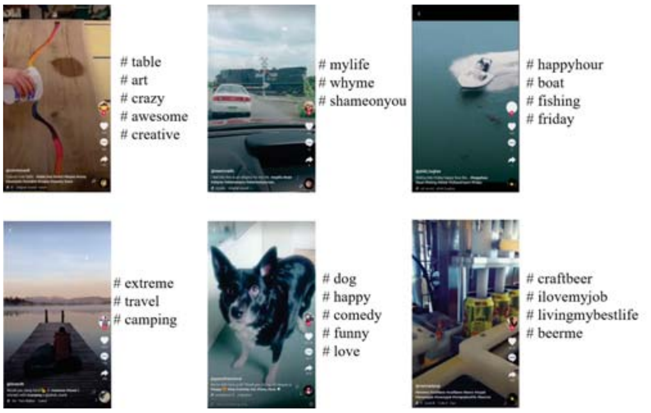


Fig. 1. Examples of hashtagged micro-videos published in Musical.ly.

1. INTRODUCTION

Since the increasing popularity of social media, it has been particularly popular and ubiquitous for people to share information online in various formats, such as text, images, videos, or any combination of above. The micro-video, characterized by various topics and short length in duration, is perhaps one of the most representative and prevalent User-Generated Content (UGC) nowadays. It exists widely in social platforms such as Musical.ly, TikTok, Instagram, etc.

A hashtag can be a single word, an unspaced phrase, or even any word combination prefixed by the symbol #, which helps users categorize and manage their posts, and easily track others' posts with a specific theme or contents. Hashtags have been employed extensively in lots of applications such as recommendation system [1], sentiment analysis [2] and topic extraction [3]. Some hashtags describe specific contents in micro-videos, while others involve information about the feelings or events which may not directly relevant to visual contents. In addition, hashtags are created by users and they hence can reflect the users' personal understanding of both micro-videos and hashtags. Ultimately, a user not only has particular bias of what kind of micro-video he/she will upload, but also has bias of what kind of hashtags to attach to the micro-video. Fig. 1 gives several examples of hashtagged micro-videos published in Musical.ly.

Hashtag recommendation aims to suggest appropriate hashtags for the given post by a specific user. These suggestions are preferably based on the micro-video contents and user personal preference. For the hashtag recommendation of text posts, Alex et al. [4] used CNNs to extract semantic embeddings from hashtags for prediction. To recommend hashtags for images, Will et al. [5] utilized statistical tagging patterns. Recently, user-conditional image hashtag recommendation [6] drew increasing attention from researchers, which considers not only item contents but also user's interest and tagging bias. As for micro-video hashtag recommendation, Yinwei et al. [7] leveraged Graph Neural Network to model the complicated interactions among users, hashtags, micro-videos to learn their representations. Since the attention mechanism has been successfully applied to extract representative regions of images [8] and important words of sentences [9], Wei et al. [10] designed a hierarchical attention network for multimodal social image popularity prediction that hierarchically attends both visual and textual information. To develop the attention model in hashtag recommendation, Zhang et al. [11] proposed a co-attention network incorporating textual and visual information to recommend hashtags for multimodal tweets. More recently, Zhang et al. [12] adopted

[*] The corresponding author.

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

a parallel co-attention mechanism to coherently model both image and text as well as the interaction between them for photo hashtag recommendation.

In this paper, we present a novel User-guided Hierarchical Multi-head Attention Network (UHMAN) for social micro-video hashtag recommendation, which learns micro-video representations conditioned by the user metadata. Specifically, the UHMAN is designed to consider characteristics of micro-video visual contents in both image level and video level and learn the hierarchical user-guided attention. In addition, we employ embedding based method to represent user conditioned micro-video contents and hashtags in embedding space, and utilize the WARP loss to optimize distances between embeddings of the micro-video and corresponding hashtags.

To the best of our knowledge, there is no publicly multimedia collection suitable for our proposed method since we take both user historical hashtags, user profile, and micro-videos into consideration. We collect a real-world dataset from a micro-video sharing platform Musical.ly to evaluate our proposed UHMAN. The experimental results demonstrate the effectiveness of the proposed method, proving the designed user-aware hierarchical attention mechanism can effectively fuse user side and micro-video side information.

The main contributions of this work are summarized as follows:

- A user-guided attention mechanism based micro-video hashtag recommendation framework is proposed, which incorporates the user profile in micro-video embedding learning effectively.

- The designed user-guided hierarchical multi-head attention mechanism includes attention for image regions and attention for video frames, which plays an important part in extracting the micro-video semantic features. WARP loss is utilized to optimize the distance between embeddings of the micro-video and its attached hashtags.

- A real-world micro-video dataset collected from Musical.ly are constructed, which consists of micro-videos, hashtags, and user profiles. Our proposed UHMAN was evaluated on the collected dataset. The experimental results demonstrate the effectiveness of our proposed UHMAN.

2. OUR PROPOSED UHMAN

2.1. Problem Definition

The hashtag recommender will suggest a rank list of hashtags for a given micro-video that a user tend to upload. Let $t \in \mathcal{T} = \{1, ..., T\}$ denote an index into a dictionary of possible hashtags. We denote the set of micro-videos tagged by the user as V. The hashtags of each micro-video is in the dictionary $\mathcal{T}$.

The UHMAN provides a sophisticated method of incorporating user metadata information when learning micro-video embeddings. The overall architecture of UHMAN is shown in Fig. 2. Intuitively, most of time users only focus on some specific regions of frames and some specific frames of the whole video when they view micro-videos. We utilize a hierarchical multi-head attention mechanism over image regions and video frames to extract statistic and dynamic semantic features of micro-video contents more effectively. The image-level attention emphasizes the specific regions, and then the video-level attention highlights frames focused by users with different demographics. Finally, hashtags are recommended based on the distances between the micro-video embedding and hashtag embeddings in the latent space.

2.2. Construction of Input Representation

Extracting visual features: The micro-video visual descriptor is obtained by a pre-trained Inception-v3 model [13]. We set frames at 5-frames-per-second and rescale all frames to 299×299 to satisfy the requirement of the input size for the model. We regard the last mixed layer of Inception-v3 as a feature extractor to gain the visual feature for each frame of dimension $8 \times 8 \times 2048$. We then reshape the visual feature to dimension 64×2048 and use a two-layer feedforward network to convert the dimension of the feature to $64 \times K$, where K denotes the hidden size. For each frame i, we get $R^i = [r_1^i, ...r_M^i]$ where $r_m^i \in \mathbb{R}^K$. M denotes the number of image regions which is equal to 64 in this work. Consequently, a micro-video can be expressed as n_f frames, each of which is expressed as 64 vectors that have dimension of K.

Encoding user representations: The user u publishing the micro-video v is originally expressed by the embedding of the user's metadata. Each user has four pieces of metadata: gender, age, country and his/her historical hashtags. Each of these information can be encoded as a one-hot encoder vector. We then apply a feedforward network to map these four metadata vectors to dimension K. Then we concatenate these representations to form the user representation u. With a feedforward network behind, we reduce the dimension of the user representation to K as the user embedding vector u_e.

Encoding hashtag representations: We define a hashtag embedding matrix W_T and perform the transformation: $t^1 = \boldsymbol{W}_T^1 t$ to convert hashtag into a low-dimensional embedding vector t^1 where $t^1 \in \mathcal{R}^K$.

2.3. User-guided Hierarchical Multi-head Attention Mechanism

The UHMAN performs user-guided multi-head attention computation in different levels, which could learn more effective video statistic and dynamic features conditioned by user

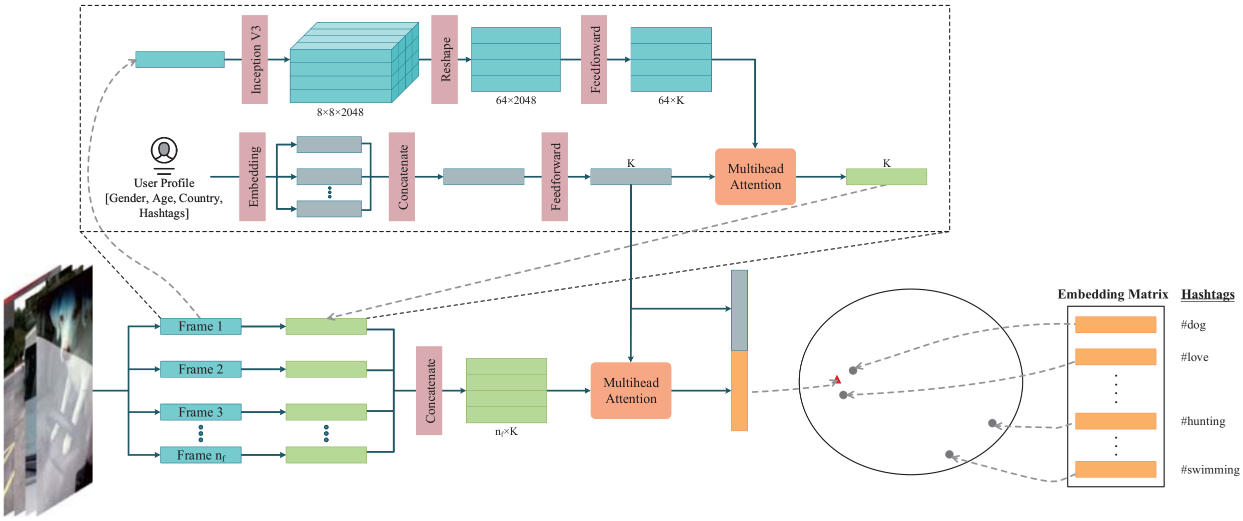


Fig. 2. The architecture of our proposed UHMAN for user conditional micro-video hashtag recommendation.

metadata. We first explicitly point out the dimension of all the input to the user-guided hierarchical multi-head attention computation, i.e., $V \in \mathbb{R}^{n_f \times 64 \times K}$, $t_e \in \mathbb{R}^K$, $u_e \in \mathbb{R}^K$. n_f denotes the number of sampled frames from a micro-video. K is the dimension of hidden size for embedding vectors.

(1) **Attention for image regions.** The goal of image level attention is to capture different importances of different image regions. The intuition lies in the aspect that users' tagging patterns are mainly related to some small specific regions of an image. Following [14], each frame extracted from a micro-video consists of 64 regional feature representations $F = \{r_1, ..., r_{64}\}$, which are fed forward to learn the user conditional image level representation $F^e = E(F)$:

$$E(F) = \Phi(\mathrm{FFN}(\Gamma(F)), \Gamma(F)) \tag{1}$$

$$\Gamma(F) = \begin{pmatrix} \Phi(MultiHead(u_e, F, F), u^e)^T \\ \ldots \\ \Phi(MultiHead(u_e, F, F), u^e)^T \end{pmatrix} \tag{2}$$

$$\Phi(\alpha, \beta) = \mathrm{LayerNorm}(\alpha + \beta) \tag{3}$$

$$\mathrm{FFN}(\gamma) = W_2 \mathrm{max}(0, W_1\gamma + b_1) + b_2 \tag{4}$$

where $\Phi(\cdot)$ represents the function that performs layer normalization on the residual output. u_e denotes the embedding representation for the user. $\Gamma(\cdot)$ denotes the multi-head attention module, and $\mathrm{FFN}(\cdot)$ is a 2-layer feedforward network, W_1, W_2 are weights for each layer, and b_1, b_2 are biases.

(2) **Attention for micro-video frames.** Based on above image level attention, we can obtain $V = [\mathrm{F}_1, ..., \mathrm{F}_{n_f}]$ for each micro-video, where $\mathrm{F}_n \in \mathcal{R}^K$ indicates the representation of one sampled frame. The goal of the video level attention is to assign frames attention weights guided by the user embedding representation, and then apply the weighted sum to construct the micro-video embedding representation. Similarly, we learn the video level representation $V^e = E(V)$ with multi-head attention module as follows:

$$E(V) = \Phi(\mathrm{FFN}(\Gamma(V)), \Gamma(V)) \tag{5}$$

$$\Gamma(F) = \begin{pmatrix} \Phi(MultiHead(u_e, V, V), u^e)^T \\ \ldots \\ \Phi(MultiHead(u_e, V, V), u^e)^T \end{pmatrix} \tag{6}$$

$$\Phi(\alpha, \beta) = \mathrm{LayerNorm}(\alpha + \beta) \tag{7}$$

$$\mathrm{FFN}(\gamma) = W_4 \mathrm{max}(0, W_3\gamma + b_3) + b_4 \tag{8}$$

After that, we calculate similarity scores between micro-video and hashtag representations in embedding space with a simple multiplication as follow:

$$f(v^e, t) = -(\boldsymbol{W}_V^2 v^e 1)^T (\boldsymbol{W}_T^2 t^1) \tag{9}$$

where $\boldsymbol{W}_V^2$, $\boldsymbol{W}_T^2 \in \mathcal{R}^{K \times d}$ and d denotes the dimension of final embedding narrow space. We concatenate the user representation u^e and the micro-video representation v^e on the first dimension to integrate user and micro-video information as the WARP input:

$$e^{uv} = Concat^0(u^e + v^e), \tag{10}$$

2.4. Training Algorithm

We train our model by minimizing the Weighted Approximate Rank Pairwise (WARP) loss described in [15], which is ideal for the hashtag recommendation task since it can easily scale to large hashtag vocabularies. In every iteration, the algorithm proceeds in the following way:

1. Sample a positive example (e^{uv}, t^{+}).
2. For the chosen (e^{uv}, t^{+}), sample a negative hashtag t^{-} such that $f(e^{uv}, t^{+}) + m > f(e^{uv}, t^{-})$, where $m > 0$ specifies the margin.
3. Make a gradient step to minimize $|m + f(e^{uv}, t^{+}) - f(e^{uv}, t^{-})|$.

Following [4], we set $m = 0.1$ in all our experiments. We uniformly sample negative hashtags t^{-} according to the fixed sampling ratio 4:1 to the number of positive hashtags in each iteration. We minimize the loss with stochastic gradient descent (SGD). The learning rate is initialized to 0.1 and is manually reduced by a factor of 10 when performance stopping improving.

3. EXPERIMENT

3.1. Datasets

To our knowledge, there is no publicly available social micro-video dataset that contains both micro-video contents and the user side information including user profile and historical hashtags. We establish such a dataset collected from Musical.ly. We remove some low frequency hashtags which have less than 10 interaction records with micro-videos. The ultimate dataset contains 10,291 micro-videos uploaded by around 6,500 de-identified users. The dataset contains 669 unique hashtags. The mean number of hashtags for per micro-video is 3.42. The de-identified users have 3 kinds of metadata: gender, age, country, and user's historical hashtags are recorded as well.

In our experiments, we used two different versions of the dataset: *natural* and *balanced*. In the *natural* dataset, we use the natural hashtag distributions of the raw dataset directly, as shown in red in Fig. 3. However, the uneven distribution of dataset labels can make infrequently used hashtags difficult to predict. To solve this problem, we created a *balanced* version of the dataset, which downsampled the 150 most common hashtags to make them similar in frequency to other hashtags. The distribution of the hashtags in blue in Fig. 3 is very close to uniform.

To evaluate the performance of the UHMAN and other baselines, we randomly split the two datasets by treating 80% micro-videos as our training set, and the remaining 20% micro-videos as our test set.

3.2. Performance and analysis

We evaluate the performance of our proposed UHMAN against the following state-of-the-art hashtag recommendation approaches:

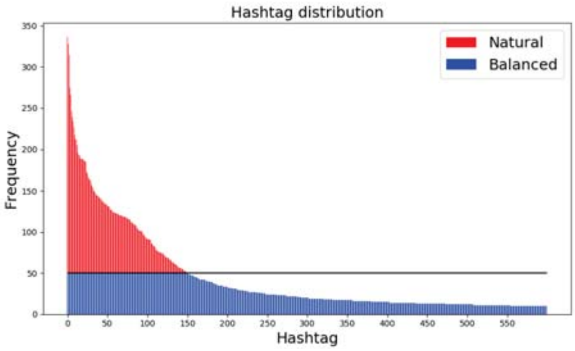


Fig. 3. Natural (red) and balanced (blue) hashtag distributions for datasets used in our experiments

- **Bilinear (BL)**. It is implemented following the similar work [16], in which we use bilinear product as similarity metric of video and hashtag embeddings for video tag prediction.
- **User_additive (UA)** [17]. It is a simple and straightforward method to incorporate user metadata information representation for micro-video embedding learning, which is also a baseline method designed for image hashtag prediction.
- **3-way-mult (3WM)** [17]. It is proposed for image hashtag prediction through a 3-way multiplicative gating visual embedding. We replace the image descriptor with the micro-video one.
- **UHMAN**. It is the method proposed by this work. A User-guided Hierarchical Multi-head Attention Network (UHMAN) with two novel hierarchical user-guided multi-head attention mechanisms to attend both image level and micro-video level modalities for learning micro-video embeddings.

The micro-video hashtag recommendation performance of all compared methods with P@10, R@10, and F1@10 on the natural version of the dataset is shown in Table 1. Table 2 demonstrates results of training on the balanced version of the dataset. From these results, we have the following observations:

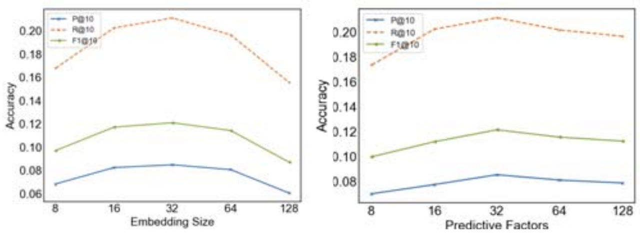
Firstly, we can see that BL has the worst performance because it only uses micro-video contents and does not consider the user side information for user's tagging bias learning. This also shows the effectiveness of incorporating user metadata information for personalized hashtag recommendation. 3WM improves Precision, Recall and F1-score compared to UA which connects user embeddings and micro-video representations directly. This also proves that combining the user information and micro-video contents effectively has better performance for personalized hashtag recommendation.

Table 1. Recommendation performance in terms of P@k, R@k, and F1@k on dataset with a natural hashtag distribution.

		P@10	P@50	R@10	R@50	F1@10	F1@50
(a)	BL	0.0650	0.0227	0.1658	0.4255	0.0934	0.0431
(b)	UA	0.0733	0.0301	0.1823	0.4389	0.1046	0.0563
(c)	3WM	0.0796	0.0312	0.1946	0.4468	0.113	0.0583
(d)	UHMAN	**0.0849**	**0.0409**	**0.2110**	**0.4517**	**0.1210**	**0.075**
Improvement	(d) vs (c)	6.66%	31.09%	8.43%	1.10%	7.08%	28.64%

Table 2. Recommendation performance in terms of P@k, R@k, and F1@k on dataset of a balanced hashtag distribution.

		P@10	P@50	R@10	R@50	F1@10	F1@50
(a)	BL	0.0371	0.0206	0.1549	0.4333	0.0599	0.0394
(b)	UA	0.0389	0.0221	0.1607	0.4604	0.0626	0.0421
(c)	3WM	0.0451	0.0237	0.1898	0.4917	0.0729	0.0453
(d)	UHMAN	**0.0507**	**0.0250**	**0.2088**	**0.5186**	**0.0816**	**0.0477**
Improvement	(d) vs (c)	12.34%	5.31%	10.00%	5.47%	11.90%	5.27%



(a)

(b)

Fig. 4. The parameter sensitivity study about the effect of (a) the embedding size K and (b) the predictive factor d. The performance of UHMAN increases as d increases to 32, and stays relatively stable as d varies.

Secondly, we observe that our proposed UHMAN outperforms other baselines (i.e. BL, UA and 3WM), which are recent competitive hashtag recommendation methods. On average, our method outperforms the second-best method by 9.5% in term of P@10, 9.22% in R@10, and 9.49% in F1@10 on the two Musical.ly datasets. This demonstrates that learning hierarchical user-guided multi-head attention for combining user metadata and micro-video contents improves the performance of user conditional hashtag recommendation.

Finally, we can see that our proposed UHMAN has consistent improvement on both datasets with natural and balance distributions. As the flatter hashtag distribution on the balanced dataset has higher entropy relative to the natural dataset, recommendation task becomes harder as evidenced by the significant reduction in all baseline's performance on P@10, R@10 and F1@10. However, our model maintains the relative performance improvements on these two datasets in general.

3.3. Parameter Sensitivity Study

We study the effects of important parameters in our method, i.e. the latent feature embedding dimension K and the dimension of the final embedding narrow space which we term it as the predictive factor d, which directly affect the final embeddings of micro-videos and hashtags. The results are shown in Fig. 4. We only show the results on the natural dataset since results on the balanced dataset show the same trend.

For the embedding size K, we vary it from 8 to 128. For simplicity, we only report the results of Precision@10, Recall@10 and F1-score@10 while omitting results of k in top-k taking other values since they show the same trend. As we can see from Fig. 4(a), with the increase of the embedding size, the recommended accuracy of UHMAN first increases and then decreases, mainly because the over-fitting occurs when the embedding size increases too much. We choose dimension 32 as a good compromise between performance and computational cost.

For the predictive factor d, which indicates the dimension of the final embedding narrow space to measure the distances between micro-video and hashtags embeddings. We vary it from 8 to 128. Similarity, we only report the results when $k = 10$ of top-k in Fig. 4(b). As we can see, the performance remains stable as d varies from 16 to 128. In our experiments, we fix the predictive factor as $d = 32$.

3.4. Ablation Study

We remove two major components of UHMAN to test their contributions to final prediction respectively. They are: a) user-guided multi-head attention for image regions, which is replaced with average pooling on image regions. b) user-guided multi-head attention for micro-video frames, which is replaced with average pooling on frames.

Table 3. Contributions of different components of UHMAN.

Methods	P@10	R@10	F1@10
UHMAN(a)	0.0736	0.1973	0.1072
UHMAN(b)	0.0822	0.2031	0.1170
UHMAN	**0.0849**	**0.2110**	**0.1210**

Table 3 shows the corresponding results on the natural dataset. The original UHMAN outperforms two aforementioned variants, and the notable performance gap further indicates the benefit of the proposed attention mechanism. Similar results are observed when k of top-k is set to other values. Based on these results, we can see the positive contribution of each component.

4. CONCLUSIONS

In this paper, a User-guided Hierarchical Multi-head Attention Network (UHMAN) is proposed for micro-video hashtag recommendation, which considers both micro-video contents and user preference. In particular, UHMAN incorporates de-identified user information, including user profile and historical hashtags, when learning micro-video embeddings. UHMAN views user information as an input query and learn both image-level and video-level attention for micro-video representations. We conduct evaluation experiments on the real-world micro-video dataset collected from Musical.ly and demonstrate the effectiveness of our proposed UHMAN.

5. ACKNOWLEDGEMENTS

This work was supported in part by National Key R&D Program of China under contract No. 2017YFB1002202.

6. REFERENCES

[1] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu, “User-video co-attention network for personalized micro-video recommendation,” in *WWW*. ACM, 2019, pp. 3020–3026.

[2] Maofu Liu, Weili Guan, Jie Yan, and Huijun Hu, “Correlation identification in multimodal weibo via back propagation neural network with genetic algorithm,” *Journal of Visual Communication and Image Representation*, vol. 60, pp. 312–318, 2019.

[3] Kar Wai Lim and Wray Buntine, “Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon,” in *CIKM*. ACM, 2014, pp. 1319–1328.

[4] Jason Weston, Sumit Chopra, and Keith Adams, “# tagspace: Semantic embeddings from hashtags,” in *EMNLP*, 2014, pp. 1822–1827.

[5] Börkur Sigurbjörnsson and Roelof Van Zwol, “Flickr tag recommendation based on collective knowledge,” in *WWW*, 2008, pp. 327–336.

[6] Hanh TH Nguyen, Martin Wistuba, and Lars Schmidt-Thieme, “Personalized tag recommendation for images using deep transfer learning,” in *ECML-PKDD*. Springer, 2017, pp. 705–720.

[7] Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie, “Personalized hashtag recommendation for micro-videos,” in *ACM MM*, 2019, pp. 1446–1454.

[8] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., “Recurrent models of visual attention,” in *NeurIPS*, 2014, pp. 2204–2212.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” in *ICLR*, 2015.

[10] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha, “User-guided hierarchical attention network for multi-modal social image popularity prediction,” in *WWW*, 2018, pp. 1277–1286.

[11] Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong, “Hashtag recommendation for multimodal microblog using co-attention network,” in *IJCAI*, 2017, pp. 3420–3426.

[12] Suwei Zhang, Yuan Yao, Feng Xu, Hanghang Tong, Xiaohui Yan, and Jian Lu, “Hashtag recommendation for photo sharing services,” in *AAAI*, 2019.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, “Going deeper with convolutions,” in *CVPR*, 2015, pp. 1–9.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017, pp. 5998–6008.

[15] Jason Weston, Samy Bengio, and Nicolas Usunier, “Wsabie: Scaling up to large vocabulary image annotation,” in *IJCAI*, 2011.

[16] Joonseok Lee, Sami Abu-El-Haija, Balakrishnan Varadarajan, and Apostol Paul Natsev, “Collaborative deep metric learning for video understanding,” in *SIGKDD*. ACM, 2018, pp. 481–490.

[17] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus, “User conditional hashtag prediction for images,” in *SIGKDD*. ACM, 2015, pp. 1731–1740.