Attentional Image Retweet Modeling via Multi-Faceted Ranking Network Learning

Zhou Zhao[1], **Lingtao Meng**[1], **Jun Xiao**[1*], **Min Yang**[2], **Fei Wu**[1], **Deng Cai**[3], **Xiaofei He**[3] and **Yueting Zhuang**[1]
[1] College of Computer Science, Zhejiang University
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] State Key Lab of CAD&CG, Zhejiang University
{csezhaozhou,lqq119119}@163.com, {junx,wufei,yzhuang}@cs.zju.edu.cn,
myang@cs.hku.hk, {dengcai,xiaofeihe}@gmail.com

Abstract

Retweet prediction is a challenging problem in social media sites (SMS). In this paper, we study the problem of image retweet prediction in social media, which predicts the image sharing behavior that the user reposts the image tweets from their followees. Unlike previous studies, we learn user preference ranking model from their past retweeted image tweets in SMS. We first propose heterogeneous image retweet modeling network (IRM) that exploits users' past retweeted image tweets with associated contexts, their following relations in SMS and preference of their followees. We then develop a novel attentional multi-faceted ranking network learning framework with multi-modal neural networks for the proposed heterogenous IRM network to learn the joint image tweet representations and user preference representations for prediction task. The extensive experiments on a large-scale dataset from Twitter site shows that our method achieves better performance than other state-of-the-art solutions to the problem.

1 Introduction

Microblog services like Twitter have become important social platforms for users to share their media contents. Retweet function is usually considered to be key mechanism that enables users to repost someone else's tweets [Zhang *et al.*, 2015b]. In social media sites, users who follows other users are termed as "followers" and users who are followed are termed as "followees". Central problem of retweet prediction is to model tweet sharing behavior that users repost tweets along followee-follower links so that more users are informed in SMS, which has attracted considerable attention recently in [Chen *et al.*, 2016; Firdaus *et al.*, 2016; Zhang *et al.*, 2015b; 2016; Feng and Wang, 2013].

*Corresponding author.

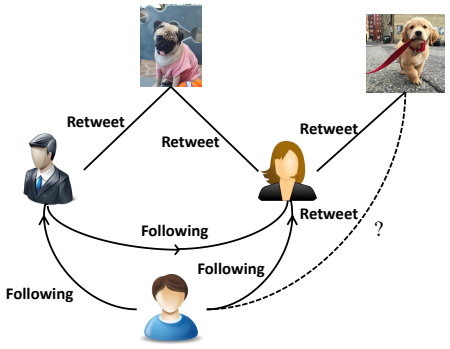


Figure 1: Example of Image Retweet Behavior.

Existing approaches for retweet prediction [Firdaus *et al.*, 2016; Zhang *et al.*, 2015b; 2016; Feng and Wang, 2013; Zhang *et al.*, 2015a] learn user preference model from their past retweeted textual tweets, and predict users' tweet sharing behavior in SMS. With the popularity of mobile devices, the amount of user-generated image tweets grows tremendously. For example, there are about 17.2% of tweets associated with images in Twitter [Chen *et al.*, 2016]. So it is important to study problem of image retweet prediction in SMS. We give a simple example of image retweet prediction in Figure 1. As there is not discriminative feature representation for tweets with image [Chen *et al.*, 2016] and SMS data is sparse [Firdaus *et al.*, 2016] , existing proposed retweet prediction methods are ineffective to image retweet prediction problem.

Currently, most of existing retweet prediction methods [Firdaus *et al.*, 2016; Zhang *et al.*, 2015b; 2016; Feng and Wang, 2013; Zhang *et al.*, 2015a] learn semantic representation of tweet based on hand-crafted feature (e.g., bag-of-words). Recently, high-level visual features for image representation with pre-trained CNNs have shown success in various visual recognition tasks [Szegedy *et al.*, 2013;

Zhao *et al.*, 2018]. Since image tweets are always visual data, it is natural to employ deep convolutional neural networks [Simonyan and Zisserman, 2014] to learn visual representation of image tweets. On the other hand, image tweets are often associated with textual context information such as users' comments and captions [Chen *et al.*, 2016]. Contextual image tweet information usually convey important messages and can gain better understanding of tweets. Since textual contextual information is always sequential data with variant length, we employ deep recurrent neural networks [Hochreiter and Schmidhuber, 1997] to learn its semantic representation. We employ multi-modal neural network learning method [Atrey *et al.*, 2010] to learn joint image tweet representation from their multi-modal contents, which provides complementary information with different modalities.

Sparsity of SMS data is also a challenging issue for image retweet prediction. In SMS sites, network between image tweets and users is constructed through users' retweet relations on image tweets. Usually, each user only retweets a few image tweets and thus SMS network is sparse. Inspired by homophily hypothesis [Yuan *et al.*, 2014], it is possible and reasonable to assume that collective information from users' followees and users' retweeted tweets can be jointly considered for tackling the sparsity problem of image retweet prediction. It is observed that social impact for retweet behavior varies between user and his/her different followees. We thus employ attention mechanism [Luong *et al.*, 2015] to adaptively incorporate users' followee preference for jointly predicting targeted user's image retweet behavior.

In this paper, we study image retweet prediction problem from viewpoint of attentional multi-faceted ranking network learning. We first propose heterogeneous image retweet modeling (IRM) network that exploits multi-modal image tweets, users' retweet behaviors and their following relations for image retweet prediction. We introduce multi-modal neural networks with two sub-networks, where recurrent neural networks learn semantic representations of image tweets' contextual information, and convolutional neural networks learn visual representations. Multi-modal fusion layer is added to learn joint image tweet representation from multi-modal neural networks. We develop attentional multi-faceted ranking method with introduced multi-modal neural networks, such that multi-faceted ranking metric is implicitly embedded in user preference representation for image retweet prediction. Main contributions of this paper are summarized as follows:

- Unlike previous studies, we present image retweet prediction problem from viewpoint of attentional multi-faceted ranking network learning. We propose heterogeneous IRM network to model the problem, which exploits multi-modal image tweets, users' retweet behaviors and their following relations.
- We develop attentional multi-faceted ranking method with multi-modal neural networks to learn user preference representation based on retweeted tweets and following relations for image tweet prediction.
- We evaluate our method's performance using dataset collected from Twitter. Extensive experiments show that our method outperforms several state-of-the-art solutions to the problem.

2 Image Retweet Prediction via Attentional Ranking Network Learning

2.1 The Problem

Before presenting the problem, we first introduce some basic notions and terminologies. Since image tweets are always visual data, it is natural to employ deep convolutional neural networks [Simonyan and Zisserman, 2014] to learn visual representation of image tweets. Given a set of image tweets $I = \{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_n\}$, we take convolutional neural networks' last hidden layer as visual representation of image tweets by $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. On the other hand, textual context information of image tweets such as users' comments and captions also gain better understanding of image tweets. We thus employ deep recurrent neural networks [Hochreiter and Schmidhuber, 1997] to learn its semantic representation. Given a set of textual contexts $D = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_n\}$, we take recurrent neural networks' last hidden layer as semantic embedding of textual contexts by $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$. We denote the joint image tweet representations by $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$, where $\mathbf{z}_i$ is joint representation of the i-th image tweet based on its visual representation $\mathbf{x}_i$ and contextual semantic representation $\mathbf{y}_i$. We denote the set of ranking models for user preference representation by $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$, where $\mathbf{u}_j$ is preference representation embedding of the j-th user.

Recently, existing approaches for retweet prediction [Firdaus *et al.*, 2016; Zhang *et al.*, 2015b; 2016; Feng and Wang, 2013; Zhang *et al.*, 2015a] learn user preference model from their past retweeted textual tweets, and then predict users' tweet sharing behavior. Unlike previous studies, we propose attentional multi-faceted ranking metric heterogeneous IRM (i.e., image retweet modeling) network that exploits multi-modal image tweets, users' past retweet behaviors and their following relations for image retweet prediction. We denote proposed heterogeneous IRM network by $G = (V, E)$, where the set of nodes V is composed of the joint image tweet representations Z and user preference representations U, the set of edges E consists of users' past retweeted behaviors H and their following relations S. We denote the retweeted behaviors between image tweets and users by matrix $H \in R^{n \times m}$, where the entry $h_{ij} = 1$ if the i-th image tweet is retweeted by the j-th user, otherwise, $h_{ij} = 0$. We then consider the following relations between users by matrix $S \in R^{m \times m}$, where $s_{ij} = 1$ if the i-th user follows the j-th user. We next denote the set of the i-th user's followees by N_i (i.e., $\mathbf{u}_j \in N_i$ if $s_{ij} = 1$), and the total set of users' followees by $N = \{N_1, N_2, \ldots, N_m\}$. We illustrate a simple example of the heterogeneous IRM network in Figure 2(a).

We then derive the heterogeneous triplet constraints from the IRM network as the users' relative preference for training the attentional multi-faceted ranking networks. We consider that the users express the explicit positive interest on the image tweets when he/she retweeted them in the IRM networks. On the other hand, following the existing Twitter analysis works [Chen *et al.*, 2012], we consider that the users

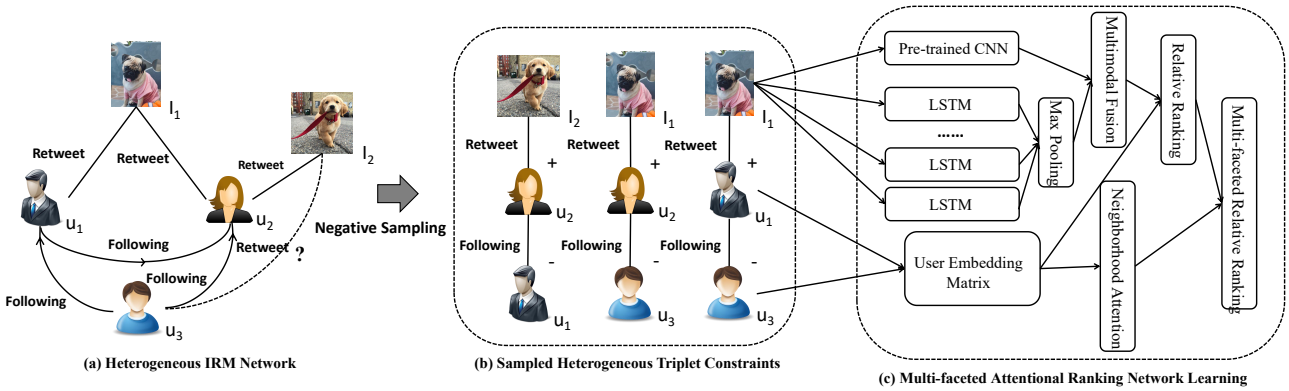


Figure 2: The Overview of Attentional Multi-faceted Ranking Network Learning for Image Retweet Prediction. (a) The heterogeneous IRM network is constructed by integrating multi-modal image tweets, users' past retweet behaviors and their following relations. (b) A negative sampling based method is employed on the heterogeneous IRM network to sample the relative users' preference. (c) The attentional multi-faceted ranking network learning method is invoked with multi-modal neural networks based on relative user preference loss for image retweet prediction.

may show the implicit negative interest on the non-retweeted image tweets of their followees. This is because the non-retweeted image tweets by the followees are more likely to be seen but disliked by the user.

Given retweeted behavior between the i-th image tweet $\mathbf{z}_i$ and the j-th user $\mathbf{u}_j$ (i.e., $h_{ij} = 1$), we sample a non-retweeted image tweet of $\mathbf{u}_j$'s followees as $\mathbf{z}_k$. Following popular homophily hypothesis [Yuan *et al.*, 2014], we also incorporate users' followee preference for image tweet modeling. We then model users' relative preference by ordered tuple (j, i, k, N_j), meaning that "the j-th user prefers the i-th image tweet to the k-th one". Let $T = \{(j, i, k, N_j)\}$ denote set of ordered tuples obtained from IRM network for a set of n image tweets and m users. We then consider ordered heterogeneous tuples as the constraints for learning user preference representations. More formally, we aim to learn the multi-faceted ranking metric function for image retweet prediction. For any $(j, i, k, N_j) \in T$, the inequality holds:

$$F_{\mathbf{u}_j}(\mathbf{z}_i) > F_{\mathbf{u}_j}(\mathbf{z}_k) \Longleftrightarrow f_{\mathbf{u}_j}(\mathbf{z}_i)h_{\mathbf{N}_j}(\mathbf{z}_i) > f_{\mathbf{u}_j}(\mathbf{z}_k)h_{\mathbf{N}_j}(\mathbf{z}_k),$$

where $F_{\mathbf{u}_j}(\cdot) = f_{\mathbf{u}_j}(\cdot)h_{\mathbf{N}_j}(\cdot)$ is the multi-faceted ranking model of the j-th user for image retweet prediction. The function $f_{\mathbf{u}_j}(\cdot)$ is the personalized ranking model of the j-th user and $h_{\mathbf{N}_j}(\cdot)$ models the social impact of the relative followee preference on the j-th user. We then define the personalized ranking function by $f_{\mathbf{u}_j}(\mathbf{z}_i) = \mathbf{u}_j^T \mathbf{z}_i$, where $\mathbf{u}_j$ is the relative preference of the j-th user and $\mathbf{z}_i$ is the joint representation of the i-th image tweet. We will present the details of the function $h_{\mathbf{N}_j}(\cdot)$ in the next section.

Using the notations above, we define the problem of image retweet prediction from the viewpoint of attentional multi-faceted ranking network learning as follows. Given the input image tweets I with their associated contexts D, the set of ordered tuples for users' relative preference T, and the heterogeneous IRM network G, our goal is to learn the multi-faceted ranking metric representations for all user preferences U and the multimodal image tweet contents Z, and then rank the image tweets for the targeted users for image retweet prediction. The image tweets to user $\mathbf{u}$ are then ranked according to the multi-faceted user preference function $F_{\mathbf{u}}(\cdot)$.

2.2 Attentional Multi-faceted Ranking Network Learning

In this section, we propose the attentional multi-faceted ranking network for image retweet prediction. We present the learning process in Figures 2(a), 2(b) and 2(c).

We first choose proper multi-modal neural networks for image tweet representation in IRM networks, which consists of two sub-networks: a deep convolutional neural network for visual representation of image data, and a deep recurrent neural network for semantic representation of textual contextual data. These two sub-networks interact with each other in a multi-modal fusion layer to form the joint representation, illustrated in Figures 2(b) and 2(c). For the visual representation of the image data, we use the activation of the 15-th layer of the proposed convolutional neural network VGGNet [Simonyan and Zisserman, 2014], which has been widely used in many visual representation tasks [Zhang *et al.*, 2017; Zhao *et al.*, 2017a; 2017b]. Meanwhile, we train the LSTM networks [Hochreiter and Schmidhuber, 1997] for the associated contexts of image tweet, and then take the output the last LSTM cell as its semantic representation. Considering the fact that the associated context of image tweets may be in the paragraph of several sentences with user comments and captions, we split them into sentences for learn the semantic representations by LSTM networks, and then fuse them by an additional max-pooling layer, shown in Figure 2(c).

In order to learn the joint representation of image tweets with different modalities, we set up the multi-modal fusion layer that connects the textual representation oriented from recurrent neural network part and visual representation oriented from convolutional neural network part, illustrated in Figure 2(c). We then map the activation of the two layers (i.e.,

the visual representation of image tweets and the semantic representation of textual contexts) into the same multi-modal feature fusion space and add them together to obtain the activation of the multi-modal fusion layer, given by

$$\mathbf{z}_i = g(\mathbf{W}^{(i)}\mathbf{x}_i + \mathbf{W}^{(d)}\mathbf{y}_i),$$

where $+$ denotes the element-wise addition for the image tweet representation with different modalities. The vector $\mathbf{z}_i$ is the joint representation of the i-th image tweet, which is obtained from the multi-modal fusion layer. The matrix $\mathbf{W}^{(i)}$ and $\mathbf{W}^{(d)}$ are weights for the fusion of multi-modal representations which can be learned in the training phrase of the attentional multi-faceted ranking network learning. The $g(\cdot)$ is the element-wise scaled hyperbolic tangent function, which forces the gradients into the most non-linear value range and leads to a faster training process, proposed in [LeCun *et al.*, 2012].

We then present the attentional multi-faceted ranking function learning for image retweet prediction. Inspired by the attention mechanism [Luong *et al.*, 2015; Zhao *et al.*, 2017c], we design the social impact function $h_{\mathbf{N}_j}(\cdot)$ based on the ordered tuple constraints $T = \{(j, i, k, N_j)\}$ as follows. Given the user preference representations $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m\}$, the social preference attention score for the p-th user and his/her q-th followee user in N_p is given by

$$s_{pq} = \mathbf{p} \cdot tanh(\mathbf{W}^{(s)}\mathbf{u}_p + \mathbf{W}^{(n)}\mathbf{u}_q + \mathbf{b}),$$

where $\mathbf{W}^{(s)}$ and $\mathbf{W}^{(n)}$ are parameter matrices to model the preference correlation between the user and his/her followee. The $\mathbf{b}$ is the bias vector and $\mathbf{p}$ is the parameter vector for computing the social preference attention score. For each followee $\mathbf{u}_q$ in N_p, its preference activation is given by $\alpha_q = \frac{\exp(s_{pq})}{\sum_{q \in N_p} \exp(s_{pq})}$. Thus, the the social impact of the relative followee preference on the j-th user is given by $h_{\mathbf{N}_j}(\mathbf{z}_i) = \sum_{\mathbf{u}_q \in N_j} \alpha_q f_{\mathbf{u}_q}(\mathbf{z}_i)$.

Given the formulation of personalized ranking function $f_{\mathbf{u}_j}(\cdot)$ and social impact function $h_{N_j}(\cdot)$, we now design the attentional multi-faceted ranking loss function as follows:

$$\mathcal{L}_{(j,i,k,N_j)} = max(0, c + F^-_{\mathbf{u}_j}(\mathbf{z}_k) - F^+_{\mathbf{u}_j}(\mathbf{z}_i)),$$

where the ranking function $F_{\mathbf{u}_j}(\mathbf{z}_i) = f_{\mathbf{u}_j}(\mathbf{z}_i) h_{\mathbf{N}_j}(\mathbf{z}_i)$, the superscript $F^+_{\mathbf{u}_j}(\cdot)$ indicates the positive preference and $F^-_{\mathbf{u}_j}(\cdot)$ denotes the negative preference. We denote the hyper-parameter c ($0 < c < 1$) controls the margin in the loss function.

We next introduce the details of our proposed attention multi-faceted ranking network learning. We denote all the model coefficients including neural network parameter, the joint image tweet representations and user preference representation by Ψ. Therefore, the objective function in our learning process is given by

$$\min_{\Psi} \mathcal{L}(\Psi) = \sum_{(j,i,k,N_j) \in T} \mathcal{L}_{(j,i,k,N_j)}(\Psi) + \beta \|\Psi\|^2,$$

where β is the trade-off parameter between the training loss and regularization term. To optimize the objective, we employ the stochastic gradient descent (SGD) with diagonal variant of AdaGrad [Kingma and Ba, 2014].

3 Experiments

3.1 Data Preparation

We collect data from Twitter, which is a popular microblog services for Web users to share their media contents [Java *et al.*, 2007]. Users usually show their positive preference on image tweets by retweeting them in social media sites. We crawl profile of the users including their past retweeted image tweets and their following relations. In total, we collect 9,900 users, 7,193 image tweets and 29,501 following relations. We report that the average time that an image tweet retweeted by some collected users is 12.2, and the average number of image tweets that some collected user retweets is 9.1. Average number of followees among the collected users is 6.2, and maximum number of followees is 162. Average number of words in the context of image tweets is 9.1, and its standard variance is 5.4. For each retweet behavior (i.e., $h_{ij} = 1$) of the user, we sample two negative image tweets from his/her followees. We sort users' retweet behaviors based on their timestamp and use the first 60%, 70% and 80% of data as training set and the remaining for testing, so the training and testing data do not have overlap. The validation data is obtained separately from the training and testing data. The dataset will be released later for further study.

3.2 Evaluation Criteria

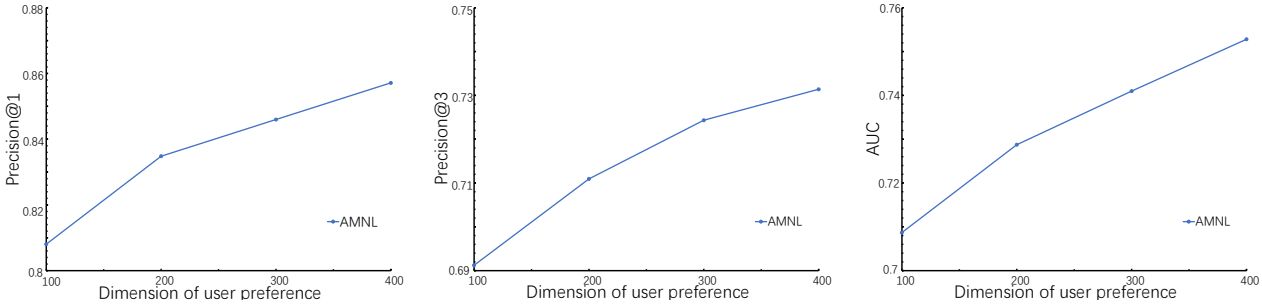
Retweet prediction task usually aims at providing top K image tweets to a user in most online media services. To evaluate the effectiveness of our method in terms of top-K ranked image tweets, we adopt two ranking-based evaluation criteria, Precision@K [Chen *et al.*, 2016] and AUC [He and McAuley, 2015; Rendle *et al.*, 2009; Li *et al.*, 2016] to evaluate the performance of image retweet prediction. Given test set of users U^t and image tweets i^t, we denote predicted ranking of the top K image tweets from test set for a certain user $\mathbf{u}_i$ by $R^{\mathbf{u}_i}$, where size of ranking list $|R^{\mathbf{u}_i}|$ is K.

3.3 Performance Comparison

We evaluate performance of our method **AMNL** with five other state-of-the-art solutions to problem of image retweet prediction: **CITING** [Chen *et al.*, 2016], **VBPR** [He and McAuley, 2015], **FAMF** [Rendle *et al.*, 2009], **ADABPR** [Rendle and Freudenthaler, 2014], **RRFM** [Li *et al.*, 2016].

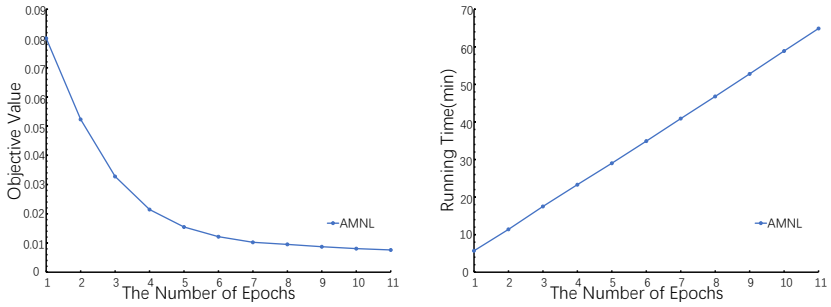
Existing retweet prediction methods are mainly based on low-rank factorized ranking model. Methods FAMF, ADABPR and RRFM learn factorized ranking metric based on pairwise preference constraints. Methods CITING and VBPR are feature-aware factorized ranking algorithms based on pairwise preference constraints and feature of item contents.

We extract feature of item contents as follows. Input words of all textual information are initialized by pre-calculated word embeddings and input visual representation of image tweets are initialized by VGG-Net [Simonyan and Zisserman, 2014]. Parameters of the neural networks used to get the representations of visual content and textual context are updated during training process. The weights of deep neural networks are randomly initialized by a Gaussian distribution with zero mean in our experiments. Following experimental setting in [Chen *et al.*, 2016; He and McAuley, 2015], we



(a) Precision@1 (b) Precision@3 (c) AUC

Figure 3: Effect of the user preference dimension on Precision@1, Precision@3 and AUC using 60% of the data for training.



(a) Objective value (b) Running time

Figure 4: Objective value and running time versus the number of epochs.

consider the associated textual contexts as the side information of the method CITING and the visual representation of image tweets as the side information of the method VBPR. The hyper-parameters and parameters which achieve the best performance on the validation set are chosen to conduct the testing evaluation. We set the learning rate to 0.01 for the gradient method. We think the top 3 tweets that users want to retweet can reveal the discriminative characteristics of the tweets that users want to retweet. So we evaluate the ranking performance of all methods on the quality of the top 3 ranked image tweets. To exploit the effect of the visual representation of image tweets and the semantic representation of the associated contexts to the performance of our method, we denote that our method with visual representation of image tweets only by AMNL$_i$, and our method with semantic representation of the associated contexts only by AMNL$_d$.

Tables 1, 2 and 3 show evaluation results of all methods on ranking criteria Precision@1, Precision@3 and AUC, respectively. Evaluation were conducted with different ratio of data as training set from 60%, 70% to 80%. We report result value of all methods using three ranking evaluation criteria. We then report performance of our model with different modalities, where dimension of user preference representation is set to 400, and 80% of data is used for training. We illustrate the effect of visual representation of image tweets, semantic representation of the associated contexts and the joint image tweet representation to our model in Table 4. These experimental results reveal a number of interesting points:

- The methods with content feature as the side information for learning the ranking metric, CITING and VBPR, outperform the low-rank factorized ranking metric methods FAMF, ADABPR and RRFM, which suggests that the deep neural networks with both image tweets and the associated context information is critical for the problem of image retweet prediction.
- Compared with other ranking methods with the side information, our method AMNL$_i$ achieves better performance than the method VBPR, and our method AMNL$_d$ achieves better performance than the method CITING, respectively. This suggests that the multi-faceted ranking metric is important for the problem.
- Compared with our methods AMNL$_i$ and AMNL$_d$, our method AMNL achieves better performance. This suggests that the attentional multi-faceted ranking network learning framework which exploits the joint image tweet representation of multi-modal image tweets and their associated context can get better performance than the attentional multi-faceted ranking network learning framework which only exploits the representation of tweets' images or the representation of tweets' contexts.
- In all cases, our AMNL method achieves the best performance. This shows that the attentional multi-faceted ranking network learning framework that exploits both the joint image tweet representation of multi-modal image tweets and their associated contexts, and multi-faceted ranking metric can further improve the performance of image retweet prediction.

In our approach, there is one essential parameter, which is the dimension of user preference representation. We vary the dimension of user preference representation from 100, 200, to 400. We show the effect of the dimension of user preference

Method	Precision@1		
	60%	70%	80%
RRFM	0.6098	0.6064	0.6261
VBPR	0.5914	0.6111	0.6215
FAMF	0.6808	0.6428	0.6071
ADABPR	0.6156	0.6134	0.6146
CITING	0.7379	0.71	0.7145
AMNL	**0.8571**	**0.8828**	**0.8604**

Table 1: Experimental results on Precision@1 with different proportions of data for training.

Method	Precision@3		
	60%	70%	80%
RRFM	0.5876	0.6188	0.632
VBPR	0.5792	0.5934	0.6252
FAMF	0.5859	0.5297	0.5066
ADABPR	0.5703	0.5903	0.6297
CITING	0.7163	0.7044	0.7391
AMNL	**0.7313**	**0.7429**	**0.7659**

Table 2: Experimental results on Precision@3 with different proportions of data for training.

Method	AUC		
	60%	70%	80%
RRFM	0.4805	0.499	0.5051
VBPR	0.5118	0.5256	0.5254
FAMF	0.5092	0.5034	0.5078
ADABPR	0.5017	0.5008	0.501
CITING	0.5004	0.5067	0.5029
AMNL	**0.7528**	**0.7977**	**0.8244**

Table 3: Experimental results on AUC with different proportions of data for training.

representation using 60% of the data for training on Precision@1, Precision@3 and AUC in Figures 3(a), 3(b) and 3(c). We find out that our method's performance trend becomes stable after the dimension of user performance presentation larger than 400 with different proportions of data for training.

The updating rule for training our proposed attentional multi-faceted ranking network learning method is essentially iterative. Here we investigate how our AMNL method converges. Figures 4(a) and 4(b) show the convergence and running time curves of AMNL method, respectively. The x-axis denotes the iteration number in both figures. The y-axis in Figure 4(a) denotes the objective value and the y-axis in Figure 4(b) shows the running time of our proposed method. Each epoch contains 22,881 iterative updates. We set the dimension of user preference representation to 400, and use 80% of the data for training. We show that our method converges after 9-th epoch and the computation cost is less than 50 minutes. This study validates the efficiency of our method.

4 Related Work

Central problem of retweet prediction is to model tweet sharing behavior that users repost tweets along followee-follower links so that more users are informed in SMS, which has attracted considerable attention recently in [Chen *et al.*, 2016; Firdaus *et al.*, 2016; Zhang *et al.*, 2015b; 2016; Wang *et al.*, 2013; Feng and Wang, 2013]. Chen et. al. [Chen *et al.*, 2016] exploit various contexts for image understanding and retweet prediction. Firdaus et. al. [Firdaus *et al.*, 2016] propose a retweet prediction model by considering user's author and retweet behaviors. Zhang et. al. [Zhang *et al.*, 2015b] propose non-parametric models to combine structural, textual, and temporal information together to predict retweet behavior. Zhang et. al. [Zhang *et al.*, 2016] propose deep neural networks to incorporate contextual and social information. Wang et. al. [Wang *et al.*, 2013] present a recommendation model to solve the problem of whom to mention in a tweet. Feng et. al. [Feng and Wang, 2013] propose the feature-aware factorization model to re-rank the tweets, which unifies the linear discriminative model and the low-rank factorization model. Peng et. al. [Peng *et al.*, 2011] model the retweet behavior using conditional random fields. Zhang et. al. [Zhang *et al.*, 2015a] employ the social influence locality for modeling the retweet behaviors. Unlike previous studies, we formulate the problem of image retweet prediction from the viewpoint of attentional multi-faceted ranking network learning, which can be solved by the negative sample based ranking metric learning with multi-modal neural networks.

Method	Precision@1	Precision@3	AUC
AMNL_i	0.8039	0.7038	0.7509
AMNL_d	0.7493	0.6749	0.7233
AMNL	**0.8604**	**0.7659**	**0.8244**

Table 4: Experimental results with different modalities using 80% of the data for training.

5 Conclusion

In this paper, we introduced problem of image retweet prediction from viewpoint of attentional multi-faceted ranking network learning. We propose heterogeneous IRM network that exploits both users' past retweeted image tweets, associated textual context and users' following relations. We present a novel attentional multi-faceted ranking network learning method with introduced multi-modal neural networks to learn joint image tweet representations and user preference representations, such that multi-faceted ranking metric is embedded in representations for prediction. We evaluate performance of our method using dataset from Twitter. Extensive experiments demonstrate that our method can achieve better performance than several state-of-the-art solutions.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant No.61602405 and No.61572431, Alibaba Innovative Research, Zhejiang province key project (2015C01027) and China Knowledge Centre for Engineering Sciences and Technology. This work is also Supported by Zhejiang Natural Science Foundation (LZ17F020001) and Key R&D Program of Zhejiang Province (2018C01006).

References

[Atrey *et al.*, 2010] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[Chen *et al.*, 2012] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. Collaborative personalized tweet recommendation. In *SIGIR*, pages 661–670. ACM, 2012.

[Chen *et al.*, 2016] Tao Chen, Xiangnan He, and Min-Yen Kan. Context-aware image tweet modelling and recommendation. In *ACM Multimedia*, pages 1018–1027. ACM, 2016.

[Feng and Wang, 2013] Wei Feng and Jianyong Wang. Retweet or not?: personalized tweet re-ranking. In *CIKM*, pages 577–586. ACM, 2013.

[Firdaus *et al.*, 2016] Syeda Nadia Firdaus, Chen Ding, and Alireza Sadeghian. Retweet prediction considering user's difference as an author and retweeter. In *ASONAM*, pages 852–859. IEEE, 2016.

[He and McAuley, 2015] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. *arXiv:1510.01784*, 2015.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Java *et al.*, 2007] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *SNA-KDD*, pages 56–65. ACM, 2007.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[LeCun *et al.*, 2012] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[Li *et al.*, 2016] Huayu Li, Richang Hong, Defu Lian, Zhiang Wu, Meng Wang, and Yong Ge. A relaxed ranking-based factor model for recommender system from implicit feedback. 2016.

[Luong *et al.*, 2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv:1508.04025*, 2015.

[Peng *et al.*, 2011] Huan-Kai Peng, Jiang Zhu, Dongzhen Piao, Rong Yan, and Ying Zhang. Retweet modeling using conditional random fields. In *ICDMW*, pages 336–343. IEEE, 2011.

[Rendle and Freudenthaler, 2014] Steffen Rendle and Christoph Freudenthaler. Improving pairwise learning for item recommendation from implicit feedback. In *WSDM*, pages 273–282. ACM, 2014.

[Rendle *et al.*, 2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452–461. AUAI Press, 2009.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[Szegedy *et al.*, 2013] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. In *NIPS*, pages 2553–2561, 2013.

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

[Yuan *et al.*, 2014] Zhaoquan Yuan, Jitao Sang, Changsheng Xu, and Yan Liu. A unified framework of latent feature learning in social media. *TMM*, 16(6):1624–1635, 2014.

[Zhang *et al.*, 2015a] Jing Zhang, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. Who influenced you? predicting retweet via social influence locality. *TKDD*, 9(3):25, 2015.

[Zhang *et al.*, 2015b] Qi Zhang, Yeyun Gong, Ya Guo, and Xuanjing Huang. Retweet behavior prediction using hierarchical dirichlet process. In *AAAI*, pages 403–409, 2015.

[Zhang *et al.*, 2016] Qi Zhang, Yeyun Gong, Jindou Wu, Haoran Huang, and Xuanjing Huang. Retweet prediction with attention-based deep neural network. In *CIKM*, pages 75–84. ACM, 2016.

[Zhang *et al.*, 2017] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017.

[Zhao *et al.*, 2017a] Wanqing Zhao, Ziyu Guan, Hangzai Luo, Jinye Peng, and Jianping Fan. Deep multiple instance hashing for object-based image retrieval. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3504–3510. AAAI Press, 2017.

[Zhao *et al.*, 2017b] Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical dual-level attention network learning. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1050–1058. ACM, 2017.

[Zhao *et al.*, 2017c] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, 2017.

[Zhao *et al.*, 2018] Zhou Zhao, Qifan Yang, Hanqing Lu, Tim Weninger, Deng Cai, Xiaofei He, and Yueting Zhuang. Social-aware movie recommendation via multimodal network learning. *IEEE Transactions on Multimedia*, 20(2):430–440, 2018.