User-guided Hierarchical Attention Network for Multi-modal Social Image Popularity Prediction

Wei Zhang
Shanghai Key Laboratory of Trustworthy Computing
East China Normal University
China
zhangwei.thu2011@gmail.com

Wen Wang
Shanghai Key Laboratory of Trustworthy Computing
East China Normal University
China
51164500120@stu.ecnu.edu.cn

Jun Wang
Shanghai Key Laboratory of Trustworthy Computing
East China Normal University
China
jwang@sei.ecnu.edu.cn

Hongyuan Zha
School of Computational Science and Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA
zha@cc.gatech.edu

ABSTRACT

Popularity prediction for the growing social images has opened unprecedented opportunities for wide commercial applications, such as precision advertising and recommender system. While a few studies have explored this significant task, little research has addressed its unstructured properties of both visual and textual modalities, and further considered to learn effective representation from multi-modalities for popularity prediction. To this end, we propose a model named User-guided Hierarchical Attention Network (UHAN) with two novel user-guided attention mechanisms to hierarchically attend both visual and textual modalities. It is capable of not only learning effective representation for each modality, but also fusing them to obtain an integrated multi-modal representation under the guidance of user embedding. As no benchmark dataset exists, we extend a publicly available social image dataset by adding the descriptions of images. The comprehensive experiments have demonstrated the rationality of our proposed UHAN and its better performance than several strong alternatives.

CCS CONCEPTS

• **Information systems** → **Content analysis and feature selection**; *Personalization*; • **Computing methodologies** → Neural networks;

KEYWORDS

Social Image Popularity; Multi-modal Analysis; Attention Network

ACM Reference Format:
Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided Hierarchical Attention Network for Multi-modal Social Image Popularity Prediction. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186026

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186026

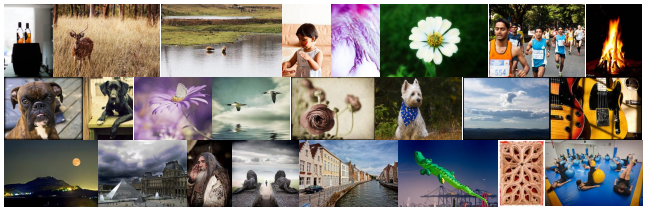

Figure 1: Sampled examples of social images in our dataset. Each row corresponds to one user. The images in each row are sorted from more popular (left) to less popular (right).

1 INTRODUCTION

In the era of Web 2.0, user-generated content (UGC) in online social networks becomes globally ubiquitous and prevalent with the development of information technology and thus incurs heavy information explosion. The task of UGC popularity prediction [35] tries to infer total count of interactions between users and specific UGC (e.g., click, like, and view). This task is crucial for both content providers and consumers, and finds a wide range of real-world applications, including online advertising [20] and recommender system [4].

Social image is perhaps one of the most representative UGC. It has gained a rapid growth in recent years and exists widely in various social medias, such as Flickr, Instagram, Pinterest, and WeChat. Due to different themes and purposes of different social medias, social images in these platforms contain not exactly the same elements. Among them, the three most common ones are social image itself (visual modality), its corresponding description (textual modality) and publisher (user). Naturally, the foregoing raises an interesting and fundamental challenge with regard to popularity prediction, i.e., how to effectively fuse knowledge from both visual and textual modalities while simultaneously consider user influence for predicting social image popularity.

While a few studies have investigated the problem of social image popularity prediction [9, 16, 40, 41], most of them largely rely on carefully designed hand-crafted features, but ignore to automatically learn joint and effective representation from multi-modalities, especially for unstructured modalities such as image and

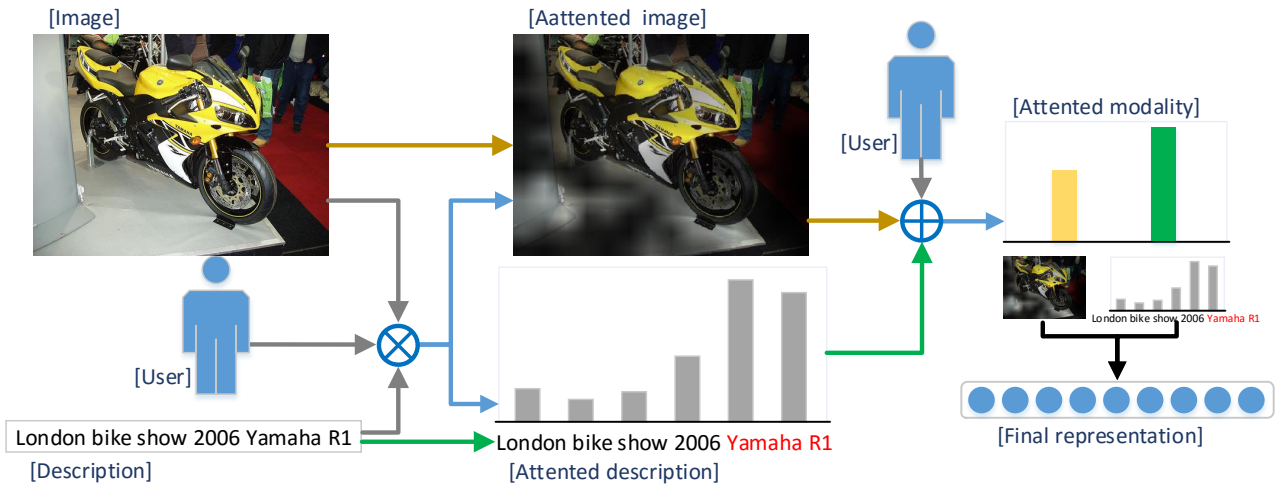

Figure 2: Diagram of user-guided hierarchical attention mechanism for an example from Flickr. ⊗ denotes the user-guided intra-attention mechanism while ⊕ represents the user-guided inter-attention mechanism. Red font in description indicates larger attention weights.

text. On the other hand, some studies have considered to combine some or all of user, text, and image information sources in their studies [7, 23, 29], and multi-modal learning has achieved great success in tasks like visual question answering (VQA) [1] and image captioning [15]. Nevertheless, the effort of applying multi-modal learning to multi-modal image popularity prediction problem has not been observed, let alone further considering user influence in multi-modal learning for this problem.

In this paper, we propose a user-guided hierarchical attention network (UHAN) for addressing the social image popularity prediction problem, which is to predict the future popularity of a new image to be published on social media. UHAN proposes two novel user-guided attention mechanisms to hierarchically attend both visual and textual modalities (see Figure 2). More specifically, the overall framework mainly consists of two attention layers which form a hierarchical attention network. In the bottom layer, the user-guided intra-attention mechanism with a personalized multi-modal embedding correlation scheme is proposed to learn effective embedding for each modality. In the middle layer, the user-guided inter-attention mechanism for cross-modal attention is developed to determine the relative importance of each modality for each user. Besides, we adopt a shortcut connection to associate the user embedding with the learned multi-modal embedding, hoping to verify its additional influence on popularity.

The intuition of utilizing user guidance behind our model is that each user has its own characteristics and preferences, which will influence the popularity of his images. To verify this, we sample several social images from three selected users and show them in Figure 1. According to the illustration below the figure, we can easily find that the user in the middle row has several images about dogs and most of them are more popular than his other images. For the user in the bottom row, a similar phenomenon can be seen that his images about cultural and natural landscapes are more attractive for ordinary users. Moreover, it is intuitive that the visual and textual modalities are promising to complement each other. This is motivated by the example shown in Figure 2, "Yamaha R1" is a major indicator for the bike in the image and vice versa. Jointly modeling them will help to capture more useful information. As there is no publicly available benchmark dataset which involves both unstructured visual and textual modalities, we build such a social image dataset by simply extending an existing publicly accessible dataset [40] by crawling their corresponding descriptions and associating them with the entries in the dataset. We conduct comprehensive experiments on this dataset and have demonstrated that 1) our proposed UHAN could achieve better results than several strong alternatives, 2) both visual and textual modalities are indeed beneficial for the studied problem, and 3) the design of UHAN is rational, with two effective user-guided attention mechanisms.

The main contributions of this work can be summarized as three-fold,

- We propose a novel user-guided hierarchical attention network that effectively learns multi-modal representation of user personalization, visual and textual modalities, and seamlessly integrates the representation learning and image popularity prediction into an end-to-end fashion.
- Two novel user-guided attention mechanisms are presented, i.e., user-guided intra-attention mechanism to learn each uni-modal representation and inter-attention mechanism to fuse multi-modal representations.
- To verify the benefits of our model, we get a real-world multi-modal social image dataset by simply extending a publicly accessible dataset [40] with crawled image title and introduction. We make the source code and the dataset[1] publicly available to facilitate other studies to repeat experiments and do further research.

[1]https://github.com/Autumn945/UHAN

2 RELATED WORK

We briefly review relevant studies to our work from three aspects. Research of popularity prediction is first introduced, including different problem settings and methods. Afterwards, deep multi-modal learning models in literature are categorized and the connection to our model is clarified. Lastly, existing representative attention mechanisms are introduced and the novelty of ours is emphasized.

2.1 Popularity Prediction

A large body of studies has focused on social media popularity prediction and this field of research has continued for more than half a decade [33, 35]. [8, 27, 37, 45] have studied social content prediction from the perspective of textual modality. Most of them are mainly based on hand-crafted features. For example, basic term frequencies and topic features extracted from topic modeling [3] are considered. By leveraging the continuous time modeling ability of point process [10], Zhao et al. [45] proposed to model dynamic tweet popularity and later Liu et al. [42] developed a feature-based point process to predict dynamic paper citation count. However, as [12] emphasized, dynamic data of popularity are not easy to obtained, which limits its real application. Thus in this paper, we focus on predicting future popularity of new social images to be published on social media.

In recent years, visual modality has attracted increasing attention in literature [5, 16, 40, 41]. Among them, Chen et al. [5] adopted transductive learning, which needs to do model learning and prediction simultaneously and cannot be easily extended to online prediction. Since the method is proposed for predicting micro-video popularity, it is different from our task. Wu et al. [40, 41] studied social image popularity from the perspective of sequential prediction. They model temporal context (i.e., feature from other images published previously) of target image for prediction, which is in parallel to our study. [9, 16] are the most relevant study to ours. However, they relies on time-consuming feature engineering to obtain various hand-crafted visual and textual features, and the feature representation and model learning are separated into two different stages.

In this paper, we explore social image popularity prediction problem by focusing on integrating the representation learning from unstructured textual and visual modalities and popularity prediction into a unified model.

2.2 Deep Multi-modal Learning

There exists a long history of studies on multi-modal learning [39] which concentrates on learning from multiple sources with different modalities [44]. In recent years, with the flourish of deep learning methodologies [21], deep multi-modal learning models begin to catch up. As Ngiam et al. [30] summarized, deep multi-modal learning involves three types of settings: 1) multi-modal fusion, 2) cross modality learning, and 3) shared representation learning. Among them, multi-modal fusion satisfies our problem setting.

Nojavanasghari et al. [31] studied persuasiveness prediction by fusing visual, acoustic and textual features with densely connected feed-forward neural network. Lynch et al. [26] proposed to concatenate deep visual features and bag-of-words based textual feature vector for learning to rank search results. To ensure fast similarity computation, hashing-based deep multi-modal learning are also proposed [14, 38]. Moreover, deep multi-modal learning has achieved a great success in VQA, developing from early simple multi-modal fusion [1] to later more complex deep methods [17, 29]. However, to our knowledge, none of multi-modal deep learning methods has been proposed to multi-modal popularity prediction task, which motivates us to take a step towards this end.

2.3 Attention Mechanism

To select important regions from images [28] or focus more on some specific words relevant to machine translation [2], attention mechanism has been proposed and sprung up. As the motivation illustrated in Section 1, we focus more on multi-modal attention. It has two important applications, i.e., visual question answering [1] and image captioning [15]. Many standard multi-modal based methods only utilize textual representation to learn attention for visual representation [6, 25, 43], without providing attentions to textual modality. Until recently, attentions to both visual and textual modalities are proposed, like dual attention networks [29]. On the other hand, personalization is rarely considered by multi-modal attention learning methods except [7]. However, this study only utilizes a single attention mechanism to generate word sequence, which leads the methodology fundamentally different from our proposed one which proposes user-guided hierarchical attention mechanism for multi-modal popularity prediction.

3 OUR PROPOSED UHAN

The overall architecture of the proposed UHAN is presented in Figure 3. The input to UHAN is a triple each time, consisting of textual representation, visual representation, and user representation, which will be clarified later. Based on this, UHAN first exploits the proposed user-guided intra-attention to learn attended embeddings for textual and visual modalities, respectively. Moreover, UHAN adopts the novel user-guided inter-attention to judge the importance of different modalities for specific users. Through this way, it further gets an attended multi-modal representation. Besides, a shortcut connection is adopted to associate user embedding with the learned multi-modal embedding for final popularity prediction.

Before we continue to specify the model, we first formally define the multi-modal social image popularity prediction problem and provide some basic notations (Section 3.1). Then we introduce the input representation for textual and visual modalities (Section 3.2). In what follows, we address the user-guided hierarchical attention mechanism (Section 3.3). Finally, popularity generation and its learning process are illustrated (Section 3.4).

3.1 Problem Definition

Before we give the formulation of the studied problem, we first introduce some mathematical notations used later. Throughout this paper, we denote matrices by bold uppercase letters and vectors by bold lowercase letters, respectively. We first indicate social image set as $\mathcal{I}$ and its size is N. As discussed in Section 1, we focus on the three most basic elements of social images. For the i-th image instance $\mathcal{I}_i$ in the set, we denote its detailed representation as $\{\mathbf{V}^i, \mathbf{H}^i, \mathbf{u}^i\}$, where $\mathbf{V}^i$, $\mathbf{H}^i$, and $\mathbf{u}^i$ correspond to visual representation, textual representation, and user representation, respectively. When the end

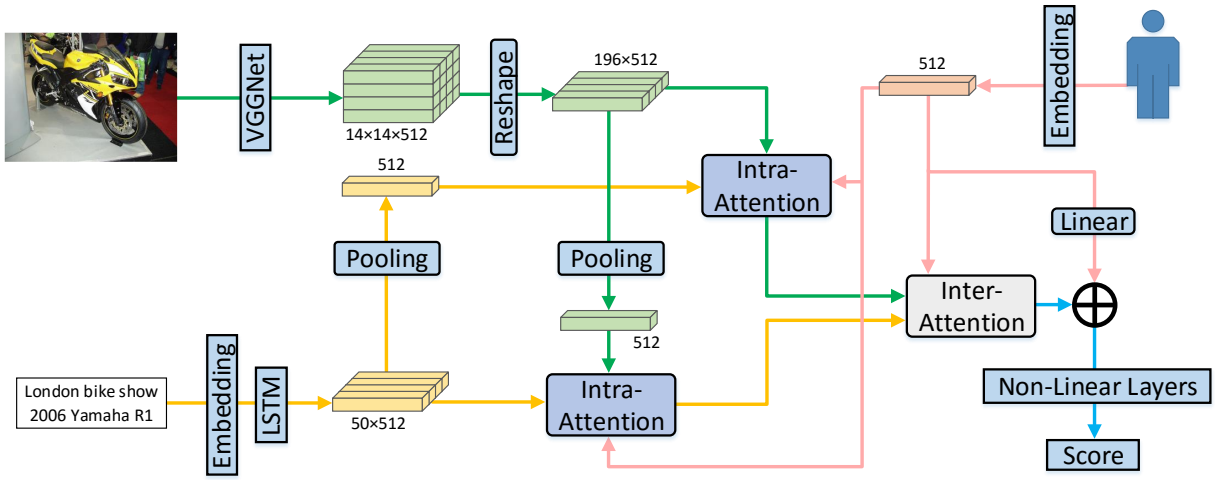

Figure 3: Architecture of our proposed model UHAN. For simplicity, the dimensions of user embedding and hidden state of LSTM are both set to 512, equal to that of visual modality. However, the above model can be easily extended to the situation that dimensions of different modalities are not equal, just by necessary linear transformation.

time is determined, we can get the real popularity score of $\mathcal{I}_i$ by considering the total number of interactions during the period of time, which is defined as y_i. Accordingly, we formally define the problem based on the above notations:

Problem 1 (**Multi-modal Social Image Popularity Prediction**). *Given a new image $\mathcal{I}_i$ to be published on social media, the target is to learn a function $f : \mathbf{V}^i, \mathbf{H}^i, \mathbf{u}^i \rightarrow y_i$ to predict its popularity score in the end.*

In what follows, we take the image instance $\mathcal{I}_i$ as an example to introduce UHAN. For simplicity, we will omit the superscript i of related notations later. In this paper, we use the terms, i.e., embedding and representation, interchangeably.

3.2 Construction of Input Representation

Extracting visual representation: The image embedding is obtained by a pre-trained VGGNet model [34]. To satisfy the requirement of the input size for the model, we first rescale all images to 448×448. By convention [29], we regard the last pooling layer of VGGNet as a feature extractor to gain visual representation $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]$ where $\mathbf{v}_m \in \mathbb{R}^{512}$. M denotes the number of image regions which is equal to 196 in this work. Consequently, an image can be expressed as 196 vectors, each of which has dimension 512.

Encoding textual representation: For the social image $\mathcal{I}_i$, it has a description $D = \{\mathbf{w}_t\}_{t=1}^{l}$ where $\mathbf{w}_t$ is a one-hot embedding at position t. l is the length of the description and should satisfy the requirement $l \leq L$, where L is the maximum length of the description and denoted as 50 in Figure 3. Hence we can get the original textual representation $\mathbf{H} = [\mathbf{w}_1, \ldots, \mathbf{w}_l]$, as required by the Problem 1.

Due to the good performance of modeling word sequence to understand language [6, 36], we further adopt long-short term memory (LSTM) [13] to encode the textual representation $\mathbf{H}$. Before we feed the one-hot embeddings of words into LSTM, we first convert each of them into a low-dimensional dense vector $\check{\mathbf{w}}_t$ by a word embedding matrix $\mathbf{W}_W$:

$$\check{\mathbf{w}}_t = \mathbf{W}_W \mathbf{w}_t. \tag{1}$$

After collecting the vectors $\{\check{\mathbf{w}}_t\}_{t=1}^{l}$, we feed them into LSTM to generate sequential hidden states. At each time step, a LSTM unit has an input gate $\mathbf{i}_t$, output gate $\mathbf{o}_t$, forget gate $\mathbf{f}_t$, and cell state $\mathbf{c}_t$. The corresponding hidden state $\mathbf{h}_t$ is calculated through the follow equations:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{Wi}\check{\mathbf{w}}_t + \mathbf{W}_{Hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \tag{2}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{Wf}\check{\mathbf{w}}_t + \mathbf{W}_{Hf}\mathbf{h}_{t-1} + \mathbf{b}_f), \tag{3}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{Wo}\check{\mathbf{w}}_t + \mathbf{W}_{Ho}\mathbf{h}_{t-1} + \mathbf{b}_o), \tag{4}$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tanh(\mathbf{W}_{Wc}\check{\mathbf{w}}_t + \mathbf{W}_{Hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \tag{5}$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t), \tag{6}$$

where $\circ$ is the Hadamard product. $\mathbf{W}_{W.}$, $\mathbf{W}_{H.}$ and $\mathbf{b}_.$ are the parameters of LSTM to be learned. σ is the sigmoid activation function. After recurrent computation for each time step, we gather a series of hidden states $\{\mathbf{h}_t\}_{t=1}^{l}$. We denote them as $\check{\mathbf{H}} = [\mathbf{h}_1, \ldots, \mathbf{h}_l]$, which will be later used in the user-guided hierarchical attention computation.

Encoding user representation: The publisher (user) of the social image $\mathcal{I}_i$ is originally expressed as a one-hot representation $\mathbf{u}$. To convert it into a low-dimensional embedding $\check{\mathbf{u}}$, we define a user embedding matrix $\mathbf{W}_U$ and perform the following transformation:

$$\check{\mathbf{u}} = \mathbf{W}_U \mathbf{u}. \tag{7}$$

Intuitively, user embeddings could capture some user hidden characteristics such as preference, which will be used to guide the learning of multi-modal representation.

In summary, we have visual representation $\mathbf{V}$, textual embeddings $\check{\mathbf{H}}$, and user embedding $\check{\mathbf{u}}$ as input for the user-guided hierarchical attention computation. We should emphasize that UHAN

will learn all the above parameters together, including the user and word embedding matrices, and the parameters of LSTM.

3.3 User-guided Hierarchical Attention Mechanism

Our model UHAN performs user-guided intra-attention and inter-attention computations in different layers, which form a hierarchical attention network that could learn more suitable representations from visual and textual modalities.

User-guided intra-attention mechanism: This attention mechanism is proposed to attend each modality to obtain textual and visual embeddings, respectively. Thus, it actually contains two attention computations, one for visual modality and the other for textual modality. However, we should emphasize that the attention computation for each modality is based on a personalized multi-modal embedding correlation scheme which involves user, visual and textual embeddings simultaneously.

We first explicitly indicate the dimension of all the input to the user-guided hierarchical attention computation, i.e., $\mathbf{V} \in \mathbb{R}^{196\times 512}$, $\check{\mathbf{H}} \in \mathbb{R}^{\mathrm{L}\times \mathrm{K}_W}$, and $\check{\mathbf{u}} \in \mathbb{R}^{\mathrm{K}_U}$. K_W and K_U are the dimensions of word and user embeddings, respectively. To be consistent with what Figure 3 shows, we let $\mathrm{L} = 50$, $\mathrm{K}_W = 512$, and $\mathrm{K}_U = 512$ for ease of presentation. Before introducing how to compute the two attentions, we should clarify that the attentions for visual and textual modalities are calculated simultaneously.

(1) Attention computation for visual modality. Based on the above specification, we illustrate how to implement the embedding correlation scheme to execute attention computation for visual modality. We convert textual embedding matrix into a vector representation $\bar{\mathbf{h}}$ through the follow equation:

$$\bar{\mathbf{h}} = \frac{1}{l} \cdot \check{\mathbf{H}}\vec{\mathbf{1}}, \tag{8}$$

where $\vec{\mathbf{1}}$ is a vector with all elements to be 1. This equation can be regarded as a mean-pooling operation applied to the hidden states of the word sequence to get an integrated textual representation for attending visual modality. After that, the representations of user and text are both vectors.

We formally define the computational formula of personalized multi-modal embedding correlation scheme for determining the visual attention as follows:

$$\mathbf{r}_{V,m} = \mathbf{W}_V^1\Big(\tanh(\mathbf{W}_{Vv}^1\mathbf{v}_m) \circ \tanh(\mathbf{W}_{Vu}^1\check{\mathbf{u}}) \circ \tanh(\mathbf{W}_{Vt}^1\bar{\mathbf{h}})\Big), \tag{9}$$

where $\mathbf{r}_{V,m}$ denotes the importance score of region m in the target image. tanh is adopted to ensure values of different modalities mapped to the same narrow space, which benefits gradient based optimization algorithms [18]. The parameter matrices of intra-attention to visual modality satisfy the following requirements, i.e., $\mathbf{W}_V^1 \in \mathbb{R}^{1\times 512}$, $\mathbf{W}_{Vv}^1, \mathbf{W}_{Vu}^1$ and $\mathbf{W}_{Vt}^1 \in \mathbb{R}^{512\times 512}$. The intuitive interpretation of the above equation is that it could be regarded as calculating the relevance of each visual region to user and textual embeddings jointly. Therefore, user and text can guide attention learning of visual modality and indicate which region of image is important to reveal popularity. Suppose α_V denotes the probability distribution of attention importance, which is given by:

$$\alpha_V = \mathrm{Softmax}(\mathbf{r}_V). \tag{10}$$

Finally, based on the attention distribution, we can gain an attended whole image representation $\dot{\mathbf{v}}$ by:

$$\dot{\mathbf{v}} = \sum_m \alpha_{V,m} \cdot \mathbf{v}_m. \tag{11}$$

(2) Attention computation for textual modality. Following Equation 8, we first define the mean-pooling formula to get a vector representation $\bar{\mathbf{v}}$ of visual modality as follows:

$$\bar{\mathbf{v}} = \frac{1}{196} \cdot \mathbf{V}\vec{\mathbf{1}}. \tag{12}$$

Likewise, attentions to each hidden state representation of the word sequence are further calculated by:

$$\mathbf{r}_{T,t} = \mathbf{W}_T^1\Big(\tanh(\mathbf{W}_{Tt}^1\mathbf{h}_t) \circ \tanh(\mathbf{W}_{Tu}^1\check{\mathbf{u}}) \circ \tanh(\mathbf{W}_{Tv}^1\bar{\mathbf{v}})\Big), \tag{13}$$

$$\alpha_T = \mathrm{Softmax}(\mathbf{r}_T), \tag{14}$$

where the parameter matrices of intra-attention to textual modality satisfy $\mathbf{W}_T^1 \in \mathbb{R}^{1\times 512}$, $\mathbf{W}_{Tv}^1, \mathbf{W}_{Tu}^1$ and $\mathbf{W}_{Tt}^1 \in \mathbb{R}^{512\times 512}$. $\mathbf{r}_{T,t}$ represents the importance score of hidden state $\mathbf{h}_t$ and α_T denotes the probability distribution of attention importance as well. It is necessary to conduct the importance calculation since some words in a textual description, including its corresponding title, may be irrelevant to popularity and even off-topic. Consequently, we can get the attended whole text embedding $\dot{\mathbf{h}}$ via the following equation:

$$\dot{\mathbf{h}} = \sum_t \alpha_{T,t} \cdot \mathbf{h}_t. \tag{15}$$

In summary, we obtain the attended whole image embedding $\dot{\mathbf{v}}$ and text embedding $\dot{\mathbf{h}}$ through the user-guided intra-attention mechanism. We further feed these two embeddings into user-guided inter-attention computation.

User-guided inter-attention mechanism: The inter-attention mechanism is proposed to capture different importance of the studied two modalities. The intuition lies in the aspect that different users have diverse concentrations on textual and visual modalities of their posted images. And even for the same user, when he is prepared to post an image, he might focus more on different modalities in different situations. The imbalance of attention mights makes the two modalities have different influence on popularity.

We denote the attention to visual modality as a_1 and textual modality as a_2, satisfying $a_1 + a_2 = 1$. Then we define the formula to calculate a_1 and a_2 through the following equations:

$$uv = \mathbf{W}_{UVT}^2\Big(\tanh(\mathbf{W}_V^2\dot{\mathbf{v}}) \circ \tanh(\mathbf{W}_U^2\check{\mathbf{u}})\Big), \tag{16}$$

$$ut = \mathbf{W}_{UVT}^2\Big(\tanh(\mathbf{W}_T^2\dot{\mathbf{h}}) \circ \tanh(\mathbf{W}_U^2\check{\mathbf{u}})\Big), \tag{17}$$

$$a_1 = \frac{\exp(uv)}{\exp(uv) + \exp(ut)}, \tag{18}$$

$$a_2 = \frac{\exp(ut)}{\exp(uv) + \exp(ut)}, \tag{19}$$

where uv denotes the relevance score between user and visual modality, and ut corresponds to user and textual modality. The parameter matrices of inter-attention computation satisfy $\mathbf{W}_{UVT}^2 \in \mathbb{R}^{1\times 512}$, $\mathbf{W}_U^2, \mathbf{W}_V^2$ and $\mathbf{W}_T^2 \in \mathbb{R}^{512\times 512}$. Upon this, we can compute the attended multi-modal embedding $\mathbf{s}$ as follows:

$$\mathbf{s} = a_1 \cdot \dot{\mathbf{v}} + a_2 \cdot \dot{\mathbf{h}}. \tag{20}$$

3.4 Learning for Popularity Prediction

To test whether the user embedding $\check{\mathbf{u}}$ has additional influence on popularity besides its major role of guiding the computation of attention to multi-modalities, we adopt a shortcut connection strategy [11] and calculate the updated multi-modal embedding as follows:

$$\mathbf{s} := \mathbf{s} + \mathbf{W}_U^3 \check{\mathbf{u}}, \tag{21}$$

where $\mathbf{W}_U^3 \in \mathbb{R}^{512\times512}$. After that, we utilize a simple 2-layer feed-forward neural network to generate final popularity prediction, which does not incur much model complexity and ensures the capacity of nonlinear modeling. More specifically, we define the computational formula as follows:

$$\hat{y} = \mathbf{W}_F^2 \text{ReLU}(\mathbf{W}_F^1 \mathbf{s} + \mathbf{b}_F^1) + b^2, \tag{22}$$

where ReLU represents the rectified linear unit, which is the nonlinear activation function with the form, ReLU(x) = max(0, x). $\mathbf{W}_F^1 \in \mathbb{R}^{512\times512}$ and $\mathbf{b}_F^1 \in \mathbb{R}^{512}$ are the parameters of the first layer. $\mathbf{W}_F^2 \in \mathbb{R}^{512}$ and $b^2 \in \mathbb{R}$ are the second layer's parameters. And $\hat{y}$ indicates the predicted popularity score we strive to generate.

We regard the learning of UHAN as a regression task. Mean square error (MSE) is adopted as the optimization metric. It is worth noting that the main focus of this paper is to consider how to effectively learn representation from unstructured visual and textual modalities for social image popularity prediction. Therefore, we do not consider modeling some structured and hand-crafted features such as social clues, user and sentiment features [5, 9, 16, 27]. However, our model could be easily extended to capture different features. One simple way is to concatenate the representation of features with the final multi-modal embedding $\mathbf{s}$ obtained by our model. Actually, we find this way can further improve the performance in our local test, which we do not introduce in the experiments.

4 EXPERIMENT

In this section, we present the detailed experimental results and some further analysis to answer the following essential research questions:

- **Q1:** What are the prediction results of the proposed UHAN compared with other strong alternatives?
- **Q2:** Does the joint considering of visual and textual modalities indeed benefit the studied problem?
- **Q3:** How does each component of UHAN contribute to the prediction performance?

Keeping these questions in mind, we first provide the details of experimental setups, including the dataset, evaluation metrics, baselines, and implementation details. Afterwards, we answer the three questions in sequence. Besides, we conduct qualitative analysis by some case studies to show the intuitive sense of our proposed UHAN.

4.1 Experimental Setup

4.1.1 Dataset. To our knowledge, there is no publicly available social image dataset which contains both unstructured visual and textual modalities for popularity prediction. We build such a dataset by extending a publicly accessible dataset[2] which is collected from Flickr [40] and has only unstructured visual modality and some structured features. For each social image in the original dataset, we further crawl its corresponding title and introduction to form the unstructured textual modality.

Given this extended dataset, we conduct the following preprocessing procedures. We first remove all non-English characters, tokenize each text, and convert each word to lowercase. We further remove words with less than five occurrences in our dataset to keep them statistically significant. Afterwards, we remove images with its description less than five words, similar to the procedure adopt in [22]. Finally, we obtain the dataset in our experiment and release it along with the source code, as introduced in Section 1.

Overall, we have about 179K social images and the statistics of the dataset is summarized in Table 1. To evaluate the performance of UHAN and other adopted methods, we split the dataset in chronological order and regard the first 70% as our training dataset, which is a little more consistent with real situation than just randomly splitting. For the rest of the dataset, we randomly adopt one third as the validation dataset to determine optimal parameters and two thirds as the test dataset to report prediction performance. Note that each user in the dataset has enough images.

Table 1: Basic statistics of the dataset.

Data	Image#	Word#	User#	Time Span
Flickr179K	179,686	70,170	128	2007-2013

4.1.2 Evaluation Metrics. As the studied problem belongs to regression task, we adopt two standard metrics, i.e., mean square errors (MSE) and mean absolute errors (MAE), which are widely used in literature [24, 40]. Denote y_i to be the ground truth for record i and $\hat{y}_i$ to be the prediction value, we can calculate MSE and MAE as follows:

$$\begin{aligned} \text{MSE} &= \frac{1}{n^{te}} \sum_{i=1}^{n^{te}} (y_i - \hat{y}_i)^2, \\ \text{MAE} &= \frac{1}{n^{te}} \sum_{i=1}^{n^{te}} |y_i - \hat{y}_i|, \end{aligned} \tag{23}$$

where n^{te} is the size of test set. We adopt the popularity score y_i calculated by [40], which is given by:

$$y_i = \log_2(\frac{c_i}{d_i} + 1), \tag{24}$$

where c is the total view count of the social image i and d represents how many days it has been from the time it has been posted to the specified end time.

4.1.3 Baselines. We compare our proposed UHAN with several carefully selected alternative methods, including some strong baselines based on multi-modal learning or attention mechanism.

- **HisAve.** The first baseline is the simplest one which regards historical average popularity as prediction. It provides benchmark performance for other methods.

[2] https://github.com/social-media-prediction/MM17PredictionChallenge

- **SVR.** Based on various hand-crafted features, [16] adopts support vector regression (SVR) for social image popularity prediction but without explicitly modeling unstructured textual modality. Following this, we additionally incorporate textual features such as TF-IDF and word embedding (GloVe [32]) while keeping basic visual features such as color and deep learning based features. We have tried different combinations of feature types and report the best results.
- **DMF.** It is a deep learning approach based on multi-modal learning. We adopt a similar deep multi-modal fusion (DMF) strategy widely used in literature [1, 26] to integrate visual representation from VGG and textual representation from LSTM.
- **DualAtt.** The last strong baseline is inspired by a recent dual attention network which involves simultaneous visual and textual attentions [29]. We adapt the one-layered version of the original one to our problem setting by utilizing user representation to guide attention learning.

To ensure robust comparison, we run each model three times and report their average performance.

4.1.4 Implementation Details. For textual modality, we set the maximum length of image description to 50 by truncating longer one. The dimension of word embedding and hidden state in LSTM are both set to 512. For visual modality, as introduced in Section 3.2, the input dimension to our model is 196×512. In addition, we set the dimension of user embedding to 512 as well.

We implement our proposed UHAN based on the Keras library. Adam with default parameter setting [19] is adopted to optimize the model, with the mini-batch size of 128. We terminate the learning process with an early stopping strategy. More specifically, we test model performance on the validation dataset every 64 batches. When the best performance keeps unchanged for more than 20 iterations, the learning process will be stopped.

Table 2: Evaluation results of our proposed UHAN and other adopted baselines in terms of MSE and MAE.

Methods	MSE	MAE
HisAve	4.070	1.575
SVR	3.193	1.385
DMF	3.004	1.339
DualAtt	2.412	1.185
UHAN (w/o u)	3.050	1.347
UHAN (w/o sc)	2.283	1.139
UHAN	**2.246**	**1.130**

4.2 Model Comparison (Q1)

Table 2 shows the performance comparison between UHAN and the compared baselines in terms of MSE and MAE. First, we can see HisAve performs much worse than all the other methods. It is consistent with our expectation since it does not consider any useful information about visual and textual modalities. By comparing DMF and SVR, we find DMF performs better, showing that deep multi-modal fusion based method is promising for this task. DualAtt further improves DMN by a significant margin. It is intuitive that DualAtt is a strong baseline since we adapt it to the studied problem by performing user attention to both visual and textual modalities separately. The comparison also reveals that considering attention mechanism in multi-modal learning is beneficial.

We further verify the role of users in our proposed UHAN by providing its two simplified versions, i.e., UHAN (w/o sc) which just removes the shortcut connection and UHAN (w/o u) that completely disregards user embedding. By comparing UHAN with UHAN (w/o sc), we see slightly better improvements, which demonstrates that the user embedding mainly utilized for attention computation can also facilitate the prediction. By testing UHAN (w/o u), we can see a notable performance drop compared with UHAN. This phenomenon shows that proposing user guidance for attention learning is indeed effective.

In summary, UHAN and its variant UHAN (w/o sc) achieve the best results among all the methods, including gaining notable improvements over the strong baseline DualAtt. We could conclude that the framework is effective and behaves well among all the adopted methods, which can answer question Q1.

4.3 Modality Contribution (Q2)

We choose two representative methods (SVR (not deep) and UHA (deep)) to test whether fusing visual and textual modalities indeed promote popularity prediction. We denote visual modality as V and textual modality as T for short, respectively. Thus "(w/o V)" means removing visual modality for corresponding methods and it is similar for "(w/o T)".

Table 3: Performance test of unstructured modalities.

Methods	MSE	MAE
SVR (w/o V)	3.214	1.392
SVR (w/o T)	3.644	1.484
SVR	3.193	1.385
UHAN (w/o V)	2.321	1.151
UHAN (w/o T)	2.337	1.149
UHAN	**2.246**	**1.130**

Table 3 presents the results of modality test. We can see that for both the baseline SVR and our model UHAN, they would suffer a clear performance drop if either textual modality or visual modality is not considered. Besides, we find that the methods of "(w/o V)" behaves a little better than those of "(w/o T)", which indicates that it might be easy to acquire knowledge from textual modality than visual modality since each words have more specific meanings than pixels. Finally, the methods of jointly fusing multi-modalities achieves the best results, reflecting that the two modalities might complement each other for the studied problem. Based on the above illustration, we can answer question Q2 that joint considering of visual and textual modalities is indeed meaningful.

4.4 Ablation Study (Q3)

We consider three major components of UHAN to test their contributions to final prediction. They are: 1) user-guided intra-attention mechanism, 2) user-guided inter-attention mechanism, and 3) shortcut connection of user embedding, just as introduced in Section 4.2.

Table 4 shows the corresponding results. Each of the middle three methods removes one of the three major components. They behave

Figure 4: Attention map visualization of two examples in our test dataset. Darker regions of images mean smaller attention weights. The lighter the font color is, the smaller attention weight the word will get.

Table 4: Contribution of different components of UHAN.

Methods	MSE	MAE
UHAN (w/o intra+inter)	2.316	1.150
UHAN (w/o intra)	2.265	1.138
UHAN (w/o inter)	2.271	1.139
UHAN (w/o sc)	2.283	1.139
UHAN	**2.246**	**1.130**

nearly the same in MAE, but have different performance in terms of MSE. By comparing with them, we find that UHAN outperforms them in both metrics. We have conducted paired t-test to show the significance of UHAN over the three variants in terms of MAE and found the difference is significant. Moreover, we compare UHAN with UHAN (w/o intra+inter) and the notable performance gap further indicates the benefit of the proposed attention mechanism. Based on these results, we see the positive contribution of each component and can answer the question Q3.

4.5 Qualitative Analysis

In addition to the above quantitative analysis, we visualize some attention maps generated by our model and some other methods to qualitatively analyze the performance.

Different models for the same example: In order to intuitively verify the advantages of our proposed UHAN, especially for the user-guided hierarchical attention mechanism, we select two image instances from our test dataset and show their attention maps for the selected attention based models in Figure 4.

We can first see our model clearly gains good visual attention maps in both two examples since it concentrates more on their key elements, which is consistent with human cognition. For the variant of our model, UHAN (w/o inter), its performance is slightly worse than UHAN in the first example, but is much worse in the second. This phenomenon indicates that the user-guided inter-attention mechanism could indeed influence the attention map learned for each modality. The attention maps generated by DualAtt seem to be not good for both images.

For the textual modality, our model shows good attentions to keywords in the descriptions. However, UHAN (w/o inter) presents an unexpected attention to the preposition 'in' in the first example. For the model of DualAtt, its major attention focuses on 'muscadine vines' in the first example. Nevertheless, this phrase might not be the one we want because it does not match with the key element in the image. Besides, its attention distribution in the second example seems to be a little chaotic. To sum up, this qualitative evaluation empirically demonstrates the effectiveness of UHAN, especially for its proposed attention mechanisms.

Our model for different examples: According to the predictions generated by our model, we select two examples with good prediction results and one with bad results, and further show them in Figure 5.

We can see clearly that the two examples in the top of the figure have good results. For both of them, the corresponding attention maps are shown in the left parts. Accordingly, we can easily focus on the important elements in the images, which meets our intuition that good attention results could lead better popularity prediction

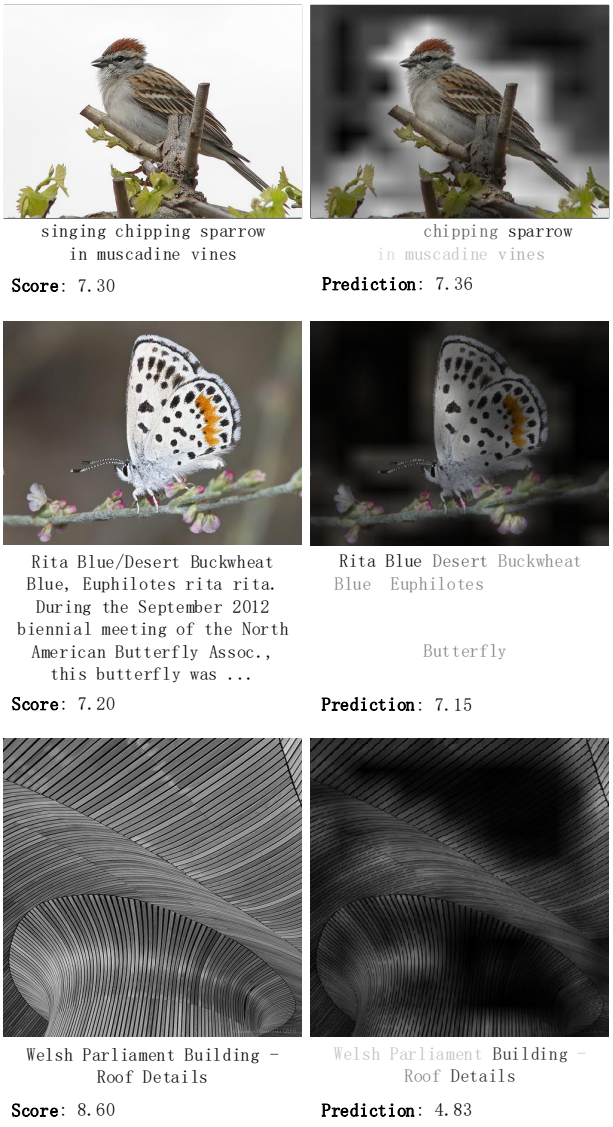

Figure 5: Case study for our model. The left column shows the original examples while the right presents the attention maps generated by our proposed UHAN and popularity scores predicted by UHAN.

performance. Moreover, by considering the last example, we find that there seems to be no obvious object or other important elements in the image. It is even not easy for ordinary users to judge its quality and popularity. Actually, some background knowledge about aesthetics might be necessary. As a result, it might be one of the main reasons that lead to an obscure attention map and poor popularity prediction result.

User personalization: In Figure 6, we select two users with different styles. "User A" usually posts images that contain people, while "User B" rarely posts this type of images, but prefers some

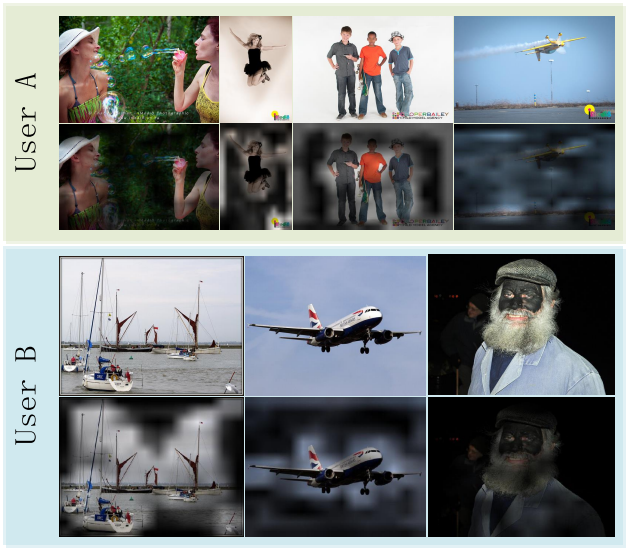

Figure 6: Case study for personalization regarding attention generation.

other objects. Therefore, we can see that attention maps generated for the images of "User A" commonly focus on people. However, for the last image of the user, it is mainly about a plane. As it does not belong to his commonly related categories, the corresponding attention map seems to be not very good as well. However, we can see that the second image of "User B" is also about a plane. But this time the generated attention map seems to be good to capture the sketch of the plane. In short, users may have different degrees of personalization, which influences attention computation and leads personalized attention maps.

5 CONCLUSION

In this paper, we have studied the problem of multi-modal social image popularity prediction. To consider representation learning from multi-modalities for popularity prediction, which is often ignored by relevant studies, we have proposed a user-guided hierarchical attention network (UHAN) model. The major novelty of UHAN is the proposed user-guided hierarchical attention mechanism that can combine the representation learning of multi-modalities and popularity prediction in an end-to-end learning framework. We have built a large-scale multi-modal social image dataset by simply extending a publicly accessible dataset. The experiments have demonstrated the rationality of our proposed UHAN and its good performance compared with several other strong baselines.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable and constructive comments. We also thank Bo Wu et al. for the released valuable dataset. This work was supported in part by NSFC (61702190), Shanghai Sailing Program (17YF1404500), SHMEC (16CG24), NSFC-Zhejiang (U1609220), and NSFC (61672231, 61672236). J. Wang is the corresponding author.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*. 2425–2433.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). arXiv:1409.0473

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[4] Biao Chang, Hengshu Zhu, Yong Ge, Enhong Chen, Hui Xiong, and Chang Tan. 2014. Predicting the Popularity of Online Serials with Autoregressive Models. In *CIKM*. 1339–1348.

[5] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model. In *MM*. 898–907.

[6] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering. *CoRR* abs/1511.05960 (2015). arXiv:1511.05960

[7] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning With Context Sequence Memory Networks. In *CVPR*. 895–903.

[8] Peng Cui, Fei Wang, Shaowei Liu, Mingdong Ou, Shiqiang Yang, and Lifeng Sun. 2011. Who should share what?: item-level social influence prediction for users and posts ranking. In *SIGIR*. 185–194.

[9] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image Popularity Prediction in Social Media Using Sentiment and Context Features. In *MM*. 907–910.

[10] Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* (1971), 83–90.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[12] Xiangnan He, Ming Gao, Min-Yen Kan, Yiqun Liu, and Kazunari Sugiyama. 2014. Predicting the popularity of web 2.0 items based on user comments. In *SIGIR*. 233–242.

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

[14] Yoonseop Kang, Saehoon Kim, and Seungjin Choi. 2012. Deep Learning to Hash with Multiple Representations. In *ICDM*. 930–935.

[15] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*. 3128–3137.

[16] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *WWW*. 867–876.

[17] Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. 2016. Multimodal Residual Learning for Visual QA. In *NIPS*. 361–369.

[18] Jin-Hwa Kim, Kyoung-Woon On, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling. *ICLR* (2017).

[19] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *ICLR*.

[20] Himabindu Lakkaraju and Jitendra Ajmera. 2011. Attention prediction on social media brand pages. In *CIKM*. 2157–2160.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[22] Kathy Lee, Ashequl Qadir, Sadid A. Hasan, Vivek V. Datla, Aaditya Prakash, Joey Liu, and Oladimeji Farri. [n. d.]. Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks. In *WWW*. 705–714.

[23] Chee Wee Leong, Rada Mihalcea, and Samer Hassan. 2010. Text Mining for Automatic Image Tagging. In *COLING*. 647–655.

[24] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. DeepCas: An End-to-end Predictor of Information Cascades. In *WWW*. 577–586.

[25] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. 2018. Co-attending Free-form Regions and Detections with Multi-modal Multiplicative Feature Embedding for Visual Question Answering. In *AAAI*.

[26] Corey Lynch, Kamelia Aryafar, and Josh Attenberg. 2016. Images Don't Lie: Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to Rank. In *SIGKDD*. 541–548.

[27] Travis Martin, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. 2016. Exploring Limits to Prediction in Complex Social Systems. In *WWW*. 683–694.

[28] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. In *NIPS*. 2204–2212.

[29] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *CVPR*. 299–307.

[30] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *ICML*. 689–696.

[31] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrusaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *ICML*. 284–288.

[32] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*. 1532–1543.

[33] Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrián, Honglin Yu, and Pascal Van Hentenryck. 2017. Expecting to be HIP: Hawkes Intensity Processes for Social Media Popularity. In *WWW*. 735–744.

[34] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556

[35] Gábor Szabó and Bernardo A. Huberman. 2010. Predicting the popularity of online content. *Journal of Commun. ACM* 53, 8 (2010), 80–88.

[36] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *ACL*. 1556–1566.

[37] Oren Tsur and Ari Rappoport. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *WSDM*. 643–652.

[38] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. 2015. Deep Multimodal Hashing with Orthogonal Regularization. In *IJCAI*. 2291–2297.

[39] William M Wells, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. 1996. Multi-modal volume registration by maximization of mutual information. *Medical image analysis* 1, 1 (1996), 35–51.

[40] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In *IJCAI*. 3062–3068.

[41] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time Matters: Multi-scale Temporalization of Social Media Popularity. In *MM*. 1336–1344.

[42] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M. Chu, and Hongyuan Zha. 2016. On Modeling and Predicting Individual Paper Citation Count over Time. In *IJCAI*. 2676–2682.

[43] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *CVPR*. 21–29.

[44] Chao Zhang, Keyang Zhang, Quan Yuan, Haoruo Peng, Yu Zheng, Tim Hanratty, Shaowen Wang, and Jiawei Han. 2017. Regions, Periods, Activities: Uncovering Urban Dynamics via Cross-Modal Representation Learning. In *WWW*. 361–370.

[45] Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. 2015. SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. In *SIGKDD*. 1513–1522.