Dynamic Attention-based Explainable Recommendation with Textual and Visual Fusion

Peng Liu[a,*], Lemei Zhang[a], Jon Atle Gulla[a]

[a] *Department of Computer Science, Norwegian University of Science and Technology, 7491, Trondheim, Norway*

Abstract

Explainable recommendation, which provides explanations about why an item is recommended, has attracted growing attention in both research and industry communities. However, most existing explainable recommendation methods cannot provide multi-model explanations consisting of both textual and visual modalities or adaptive explanations tailored for the user's dynamic preference, potentially leading to the degradation of customers' satisfaction, confidence and trust for the recommender system. On the technical side, Recurrent Neural Network (RNN) has become the most prevalent technique to model dynamic user preferences. Benefit from the natural characteristics of RNN, the hidden state is a combination of long-term dependency and short-term interest to some degrees. But it works like a black-box and the monotonic temporal dependency of RNN is not sufficient to capture the user's short-term interest.

In this paper, to deal with the above issues, we propose a novel Attentive Recurrent Neural Network (Ante-RNN) with textual and visual fusion for the dynamic explainable recommendation. Specifically, our model jointly learns image representations with textual alignment and text representations with topical attention mechanism in a parallel way. Then a novel dynamic contextual attention mechanism is incorporated into Ante-RNN for modelling the complicated correlations among recent items and strengthening the user's short-term interests. By combining the full latent visual-semantic alignments and a hybrid attention mechanism including topical and contextual attentions, Ante-RNN makes the recommendation process more transparent and explainable. Extensive experimental results on two real world datasets demonstrate the superior performance and explainability of our model.

Keywords: Dynamic Explainable Recommendation, Recurrent Neural Network, Attention Mechanism, Semantic Alignment, Multi-model Fusion, User Interests

1. Introduction

In recent years, explainable recommendation has become an active research topic in many online customer-oriented applications, such as social media, e-commerce and content-sharing websites. By explaining how the system works and/or why an item is recommended, the system becomes more transparent and has the potential to allow users to tell when the system is wrong (scrutability), help users make better (effectiveness) and faster (efficiency) decisions, convince users to try or buy (persuasiveness), or increase the ease of the user enjoyment (satisfaction) (Tintarev and Masthoff, 2011). Current explainable models usually interpret the recommendations based on user reviews. For instance, Zhang et al. (2014) proposed an Explicit Factor Model (EFM) to learn user cared features from the review information and fill them into pre-defined templates regarded as explanations. Chen et al. (2016) and Wang et al. (2018c) extended EFM for more accurate user-item-feature explanations by leveraging tensor factorization techniques. Chen et al. (2018a) used attention mechanism to extract valuable item reviews for explaining the rating prediction.

[*] Corresponding author

Email addresses: peng.liu@ntnu.no (Peng Liu), lemei.zhang@ntnu.no (Lemei Zhang), jon.atle.gulla@ntnu.no (Jon Atle Gulla)

Despite effectiveness, these explainable recommendation methods still suffer from some inherent issues: (1) Most of them model the item's characteristics by only leveraging their textual features, which leads to the limited recommendation performance and explanatory capability. In fact, for some types of items (e.g., clothing), their visual appearances play an important role in their properties, which can greatly bias the user's preference towards them. For example, users can easily determine whether they watch a movie based on the movie poster images. Thus, the visual features of items are also important complementary information for the explainable recommendation. (2) Most methods assume that user preferences are invariant and generate static explanations. However, in real scenarios, a user's preference is always dynamic, and s/he may be interested in different topics at different states. The static assumption can easily lead to incorrect matches between the explanation and user dynamic preference, thus impairing the recommendation performance and degrading customers' satisfaction, confidence and trust for the recommender system.

Previous works that leverage the visual information for personalized recommendation usually transform images into embedding vectors, which are then incorporated with collaborative filtering (CF) for improving the performance. For example, McAuley et al. (2015) adopted neural networks to transform images into feature vectors, and used the vectors for product style analysis and recommendation; He and McAuley (2016) further extended the approach to pair-wise learning to rank for recommendation; Geng et al. (2015) adopted image features for recommendation in a social network setting; Wang et al. (2017) extracted image features with neural network for point-of-interest recommendation. Though the recommendation performance has been improved by incorporating image representation extracted with (convolutional) neural networks, the related works have largely ignored an important advantage of leveraging images for recommendation – its ability to provide intuitive visual explanations. This is because by transforming the whole image into a fixed latent vector, the images become hardly understandable for users, which makes it difficult for the model to generate visual explanations to accompany certain recommendations.

On the other hand, recent approaches that leverage Recurrent Neural Network (RNN) for recommendation have demonstrated their effectiveness in modelling the temporal dynamics of user preferences. RNN based methods adopt the last hidden state as the user's final representation to make recommendations. With the help of gated activation function like long-short term memory or gated recurrent unit (Cho et al., 2014), RNN can better capture the long-term dependency. However, it works like a black-box, for which the reasons underlying a prediction cannot be explicitly presented. Besides, due to the recurrent structure and fixed transition matrices, RNN holds an assumption that temporal dependency has a monotonic change with the input time steps (Liu et al., 2017). It assumes that the current item or hidden state is more significant than the previous one. This monotonic assumption would restrict the modelling of user's short-term interests and can not well distinguish the importance of several recent factors. For example, a user is looking for interesting movies on the Internet. During browsing, s/he tends to click on some movies with the "disaster" topic which is treated as the user's short-term interest, meanwhile s/he might click a comedy movie by accident or due to curiosity. In this case, small weight should be provided for the comedy movie. So the short-term interest should be carefully examined and needs to be integrated with the long-term dependency.

In this paper, we focus on the problem of simultaneously multi-model explanation generation and dynamic user preference modelling in the context of explainable recommendation. The problem setup is illustrated in Figure 1. We propose a novel Attentive Recurrent Neural Network (Ante-RNN) to address this problem. More specifically, we first learn image representations with the latent semantic alignments between image regions and the corresponding words in text. Meanwhile, in order to capture the user's long-term preference, a topical attention mechanism which can model the interactions between the words and the user's interested topics is adopted to learn text

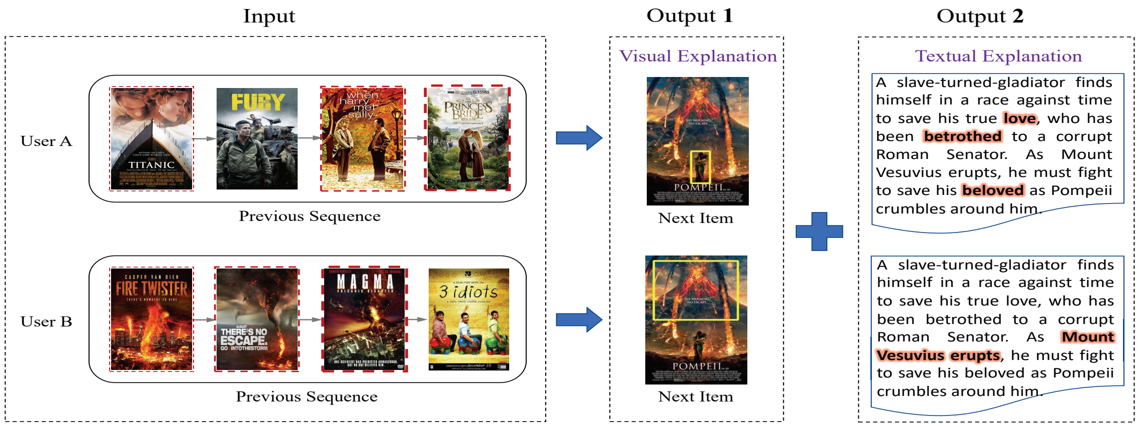

Input
Output 1
Output 2
User A
Previous Sequence
User B
Previous Sequence
Visual Explanation
Next Item
Next Item
Textual Explanation
A slave-turned-gladiator finds himself in a race against time to save his true love, who has been betrothed to a corrupt Roman Senator. As Mount Vesuvius erupts, he must fight to save his beloved as Pompeii crumbles around him.
A slave-turned-gladiator finds himself in a race against time to save his true love, who has been betrothed to a corrupt Roman Senator. As Mount Vesuvius erupts, he must fight to save his beloved as Pompeii crumbles around him.

Figure 1: Problem Setup. Given the users' clicking sequence of items, different parts of the images are marked in rectangle to provide intuitive explanations for the next recommended item. Meanwhile, their textual explanations are also provided by highlighting the topic-related words. Besides, the item in the red dashed is more relevant to the current user's intention. And the red line is thicker when the item is more important.

representations. After that, the representations from image and text sources are integrated to obtain a joint representation of visual and textual features for each item by investigating different modality fusion strategies, which is used as the input of Ante-RNN. Then a novel dynamic contextual attention mechanism is incorporated into our model for modelling the complicated correlations among recent items and strengthening the user's short-term interests. By combining the full latent visual-semantic alignments and the attention weights learned from topical attention network and contextual attention network, Ante-RNN makes the recommendation process more transparent and explainable. Compared with existing methods, our model not only improves the recommendation performance, but also generates textual and visual explanations for the recommended items.

To summarize, this paper makes the following contributions:

- We propose an Attentive Recurrent Neural Network (Ante-RNN) for the dynamic explainable recommendation which could provide multi-model explanations according to the user dynamic preference. To the best of our knowledge, it is the first time to jointly explore multi-modal and adaptive explanations in a unified framework for the personalized recommendation.

- In order to alleviate the issues caused by the monotonic assumption of RNN, a hybrid attention mechanism is developed to capture the user's long-term dynamic interest over different topics and strengthen the short-term interest simultaneously. More importantly, our proposed dynamic contextual attention scheme incorporates diverse temporal factors of the user's clicking sequence of items (e.g. time interval and the time of week) to further improve the recommendation performance.

- We analyze and study a variety of fusion strategies for mutual association learning across modalities, and find that the attention-based fusion robustly achieves the best results.

- We conduct extensive experiments on two real large-scale datasets. The results show that Ante-RNN outperforms state-of-the-art baselines in terms of Recall and NDCG on both datasets.

The remainder of the paper is organized as follows. Section 2 introduces the related work. In section 3, we formally define the problem and our new model. We describe the datasets, comparative approaches, the evaluation criteria we use and experimental results in section 4. Finally, we present the conclusions and future work in Section 5.

2. Related Work

2.1. Explainable Recommendation

Researchers have shown that providing appropriate explanations could improve user acceptance of the recommended items (Herlocker et al., 2000), as well as benefit user experience in various other aspects, including system transparency, user trust, effectiveness, efficiency, satisfaction and scrutability (Tintarev and Masthoff, 2011). However, the underlying algorithm may influence the types of explanations that can be constructed. In general, the computational complex algorithms within various latent factor models make the explanations difficult to be generated automatically (Tintarev and Masthoff, 2011). Many meticulously designed strategies have been investigated to tackle the problem. For instance, the authors in (McAuley and Leskovec, 2013) aligned user/item latent factors in matrix factorization (MF) with topical distribution in latent dirichlet allocation (LDA) for joint parameter optimization under the supervision of both score ratings and textual reviews, and thus the user preferences are explained by the learned topical distributions. Ling et al. (2014) applied topic modelling techniques with mixture of Gaussians on the reviews and generated interpretable topics. Bao et al. (2014) further extended Ling's work and proposed a novel topical matrix factorization model (TopicMF) to extract topics from each review. To explain finer-grained user preference, some approaches (Zhang et al., 2014; Chen et al., 2016) combined matrix factorization (MF) and sentiment analysis (SA) to generate explanations at the feature-level. More specifically, they extracted feature-opinion-sentiment triplets from the user review information, and infused them into MF for collective user preference modelling. The explanations were provided by filling the predicted user cared features into pre-defined templates. Despite effectiveness, the final results of these methods can be easily affected by the accuracy of the review preprocessing tools, and the complex process for extracting triplets usually render them inefficient.

Recently, with the rapid development of deep learning technology, there has been a surge of interest in leveraging attention mechanisms to explain a recommendation. A common approach employed by these works involves two steps. First, they merge the raw user review information related to a user (or an item) into a document and attentively discovered valuable information in the document. The next step is to provide the explanations by highlighting the words with the highest attention weights. In particularly, Seo et al. (2017) and Chen et al. (2018a) automatically learned the importances of different review sentences under the supervision of user-item rating information. To provide explanations tailored for different target items, Tay et al. (2018) adopted ”co-attention” mechanism to capture the correlations between users and items. Apart from user-review explanations, Ai et al. (2018) conducted explainable recommendation by reasoning over knowledge graph embeddings, where explanation paths between users and items were constructed to generate knowledge-enhanced explanations. Hu et al. (2018) built a multilevel personal filter to calculate users' attractiveness on textual information of items and provided interpretable recommendations upon them.

Although these methods have achieved promising results, they failed to model user dynamic preference, and the provided explanations were usually static and unimodal, which may weaken the persuasiveness of the explanations as mentioned before.

2.2. Sequence-aware Recommendation

Recently, a number of research works have demonstrated that the sequential information (e.g., user sequential behaviors), which are regarded as the important information source for understanding user dynamic preferences, can be utilized to improve personalized recommendations at the right time. In specific, early methods care more about transition properties between two successive behaviors. For instance, the factorized personalized Markov chains (FPMC) (Rendle et al., 2010) combined matrix factorization with one-order Markov chain to capture the influence of the last behavior towards the next one. The hierarchical representation model (HRM) (Wang et al., 2015) generalized FPMC into a representation learning framework, and significantly improved the recommendation performance. The major limitation of these methods lies in the ignoring of

Table 1: Summary of related studies about the sequence-aware recommendation. Fields without information in the related study are marked with a hyphen.

		Features			
Reference	**Model**	**Short-term Behaviors**	**Long-term Behaviors**	**Relevance of Historical Behaviors**	**Temporal Context**
Rendle et al. 2010	Markov chain	✓	-	-	-
Wang et al. 2015	Markov chain	✓	-	-	-
Hidasi et al. 2015	RNN	✓	✓	-	-
Yu et al. 2016	RNN	✓	✓	-	-
Song et al. 2016	RNN	✓	✓	-	-
Zhu et al. 2017	RNN	✓	✓	-	✓
Chen et al. 2018b	RNN + Memory network + Attention mechanism	✓	✓	✓	-
Huang et al. 2018	RNN + Memory network + Attention mechanism	✓	✓	✓	-
Wang et al. 2018b	DNN + Attention mechanism	✓	✓	✓	-
Pei et al. 2017	RNN + Attention mechanism	✓	✓	✓	-
Li et al. 2017	RNN + Attention mechanism	✓	✓	✓	-
Ante-RNN	RNN + Attention mechanism	✓	✓	✓	✓

long-term preference dependency.

To solve this problem, many models were proposed to capture user multi-step behaviors based on the recurrent neural network (RNN). Yu et al. (2016) represented a basket acquired by pooling operation as the input layer of RNN, which outperforms the state-of-the-art methods for next basket recommendation. Song et al. (2016) proposed a multi-rate Long Short-Term Memory (LSTM) with considering temporal user preferences for commercial news recommendation. Hidasi et al. (2015) utilized RNNs for session-based online recommendation. Furthermore, with the ability to express, store and manipulate the records explicitly, dynamically and effectively, external memory networks (EMN) (Sukhbaatar et al., 2015) have shown their promising performance for many sequential prediction tasks, such as question answering (QA) (Kumar et al., 2016), natural language transduction (Grefenstette et al., 2015), and recommender system (Chen et al., 2018b). Chen et al. (2018b) proposed a novel framework integrating recommender system with external User Memory Networks which could store and update users' historical records explicitly. Huang et al. (2018) proposed to extend the RNN-based sequential recommender by incorporating the knowledge-enhanced Key-Value Memory Network (KV-MN) for enhancing the representation of user preference. Despite these models achieve some degree of improvements, one of the important features - the temporal context of user sequential behaviors - has been totally ignored. Recently, Zhu et al. (2017) designed a model called Time-LSTM to demonstrate the importance of time interval information for user dynamic preference modelling. However, their proposed model was designed for a particular type of contextual information (i.e. time intervals) and is not flexible to incorporate other types of context (e.g. the time of week). What's more, the Time-LSTM model cannot automatically select important interaction records in the user-item interaction history when recommending items.

To model the different impacts of a user's diverse historical interests on current candidate item, Wang et al. (2018b) designed an attention module to dynamically calculate a user's aggregated historical representation. Pei et al. (2017) extended recurrent networks for modelling user and item dynamics with a novel gating mechanism, which adopts the attention model to measure the relevance of individual time steps of user and item history for recommendation. Li et al. (2017) explored a hybrid encoder with an attention model to capture both the user's sequential behavior and main purpose in the current session. Specifically, they involved an item-level attention mechanism which allowed the decoder to dynamically select and linearly combine different parts of the input sequence. Different from existing works, we propose a hybrid attention mechanism which takes into account the user's long-term interested topics and short-term contextual surroundings at the same time. And more importantly, the proposed dynamic contextual attention scheme enables our model to selectively concentrate on critical parts of the sequential information and is fairly flexible, which can easily add other types of contextual information when available. To illustrate more clearly, Table 1 summarizes the differences among related works with sequential information.

2.3. Multi-modality Fusion

Multi-modality fusion enables us to leverage complementary information presented in multimodal data, thus discovering the dependency of information on multiple modalities. There exist two commonly used fusion strategies in previous research: feature-level fusion and decision-level fusion. Specifically, feature-level fusion aims to directly combine feature vectors by concatenation (Sun et al., 2018) or kernel methods (Bucak et al., 2014; Poria et al., 2016). Poria et al. (2016) used a multiple kernel learning strategy to fuse the modality data on the feature-level. Zadeh et al. (2017) proposed a tensor fusion technique to fuse audio, visual and textual features at feature level. Decision-level fusion builds separate models for each modality and then integrates the outputs together using a method such as majority voting or weighted averaging (Wörtwein and Scherer, 2017; Nojavanasghari et al., 2016). For instance, Wöllmer et al. (2013) combined the results of the text and audio-visual modalities by a threshold score vector on the decision-level. Deep neural network fusion was proposed in a recent study to fuse the extracted modality-specific features (Zhang et al., 2018; Liang et al., 2018). More recent approaches introduced LSTM structures to fuse the features at each time step (Poria et al., 2017; Chen et al., 2017).

In recommender systems, previous works often adopt the strategy of combining image-, rating- and review-based features for boosting recommendation performance. The most frequently used fusion methods are concatenation (Sun et al., 2018; Guan et al., 2019), addition (Tan et al., 2016; Zhang and Wang, 2016) and element-wise product (Wang et al., 2018a). Recently, Zhang et al. (2017) integrated images with reviews and ratings in a multimodal deep learning framework for top-n recommendation. Cui et al. (2018) proposed a multi-modal Marginalized Denoising AutoEncoder (3mDAE) to learn fusion features by reconstructing the original multi-modal data. However, only few works consider the sophisticated interactions between different modalities in the recommendation. For instance, Cheng et al. (2018) adopted a fully-connected neural layer directly after the addition fusion step to get better fusion features in the rating prediction. Lian et al. (2018) proposed a multi-channel deep fusion model which leverages an attention mechanism to merge latent representations learnt from different domains in the personalized news recommendation. In this work, we explore several fusion techniques for mutual association learning across modalities (mainly based on the textual and visual modalities) in the context of explainable recommendation.

3. Proposed Ante-RNN Model

In this section, we describe the proposed Attentive Recurrent Neural Network (Ante-RNN) for the dynamic explainable recommendation in detail. The basic idea of Ante-RNN is to build a unified representation of the user's interacted items, and then generate predictions along with explanations based on it. The representation should take into account various potential factors that influence user's next decision. As shown in Figure 2, our model firstly learns text embedding with topical attention network fused with image embedding with the according textual alignment in the same D-dimensional space to represent item. Then our dynamic contextual attention mechanism learns attentive weights by considering the contextual influence of current interacting (e.g. clicking/reading) item to strengthen the representation before GRU network, and thus to improve the recommendation performance. Furthermore, the attention weights learned from topical attention network and contextual attention network, can in turn help to explain the recommenda-

Table 2: Notations used in the paper.

Symbol	Description
$\mathcal{U}, \mathcal{I}$	The set of users and items
$\mathcal{E}, \mathcal{F}$	The set of word features and image features for an item
M, N	The number of image features and word features for an item
$\boldsymbol{v}, \boldsymbol{x}$	The image embedding and the text description embedding
D	Dimensionality of the image and text embedding
$\boldsymbol{C}$	The affinity matrix whose element represents the similarity between image region and text word
$\boldsymbol{\eta}_t^u$	User u's historical interested topics representation
ψ	The number of topics
$\boldsymbol{i}$	Item representation after multi-model fusion
w_c	The contextual window length
$\tilde{\boldsymbol{r}}_t$	User's interest representation at timestamp t
$\boldsymbol{T_w}$	One hot encoding vector of time of week
$\boldsymbol{\delta t}$	One hot encoding vector of time interval
$\boldsymbol{o}_t$	The output vector from GRU

tion results by the descriptive snippets learned from images and texts.

In the rest of this section, we first define relevant notations used in this paper and formulate the recommendation problem. Then, we present the image embedding with textual alignment and topic-based text embedding in Section 3.2 and Section 3.3 respectively. In Section 3.4, several multi-model fusion strategies are explained in detail. We introduce the dynamic contextual attention mechanism in Section 3.5 and finally in Section 3.6, the whole objective of our model and its training procedure will be described.

3.1. Problem Formulation

Throughout this paper, all vectors are column vectors and are denoted by bold lower case letters (e.g. $\boldsymbol{x}$ and $\boldsymbol{y}$), while matrices are represented by bold upper case letters (e.g., $\boldsymbol{X}$ and $\boldsymbol{Y}$). We use calligraphic letters to represent sets (e.g., $\mathcal{U}$ and $\mathcal{I}$). Lower case letters (e.g. x and y) represent as scalar parameters. Table 2 summarizes the notations of frequently used variables.

Let $\mathcal{U} = \{u_i, u_2, ..., u_{|\mathcal{U}|}\}$ and $\mathcal{I} = \{i_1, i_2, ..., i_{|\mathcal{I}|}\}$ represent the sets of users and items respectively. For each user, we chronologically organize his/her historical behaviors as a sequence of tuples $O^u = \{(i_1^u, t_1^u), (i_2^u, t_2^u), ..., (i_{l_u}^u, t_{l_u}^u)\}$ with the length l_u, where $t_1^u \leq t_2^u \leq ... \leq t_{l_u}^u$, and the s-th element (i_s^u, t_s^u) means that user u interacted with (i.e. clicked/viewed) item $i_s^u \in \mathcal{I}$ at time t_s^u. Additionally, an image and a text description are available for each item $i \in \mathcal{I}$. Our task of explainable recommendation with user dynamic preference is to learn a model such that for any given user's historical interacted item set O^u, it generates a list of top-k personalized items as recommendations for user u. And further, its internal parameters or intermediate outputs should provide explanations on both textual and visual modalities for these recommended items according to the user's preference at time $t_{l_u+1}^u$.

3.2. Image Embedding with Textual Alignment

Inspired by the work of Lee et al. (2018), we learn image and its corresponding text description in a joint manner. Though items can be expressed by multiple ways such as image, video, sound, text and so on, the combined representations of items should require a feature fusion mechanism to ensure that multiple inputs are appropriately integrated. Furthermore, the strategy that synchronizes different inputs of multi-modalities at the same level is an effective way as well (Gu et al., 2018a). Therefore, in this paper, we consider to represent image at the word level because word is an important basic unit of representing users' interests, and thus image-based item representation and text/word-based item representation can be projected at the same space.

Suppose an item includes a set of word features $\mathcal{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_N\}$, in which each element $\boldsymbol{e}_i \in \mathbb{R}^D, i = 1, 2, ..., N$ denotes a word representation in the text description, and a set of image features $\mathcal{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_M\}$, in which each element $\boldsymbol{f}_j \in \mathbb{R}^D, j = 1, 2, ..., M$ denotes a region representation in the image. Same with Lee et al. (2018), the image region representations are derived by adopting the Faster R-CNN model in conjunc-

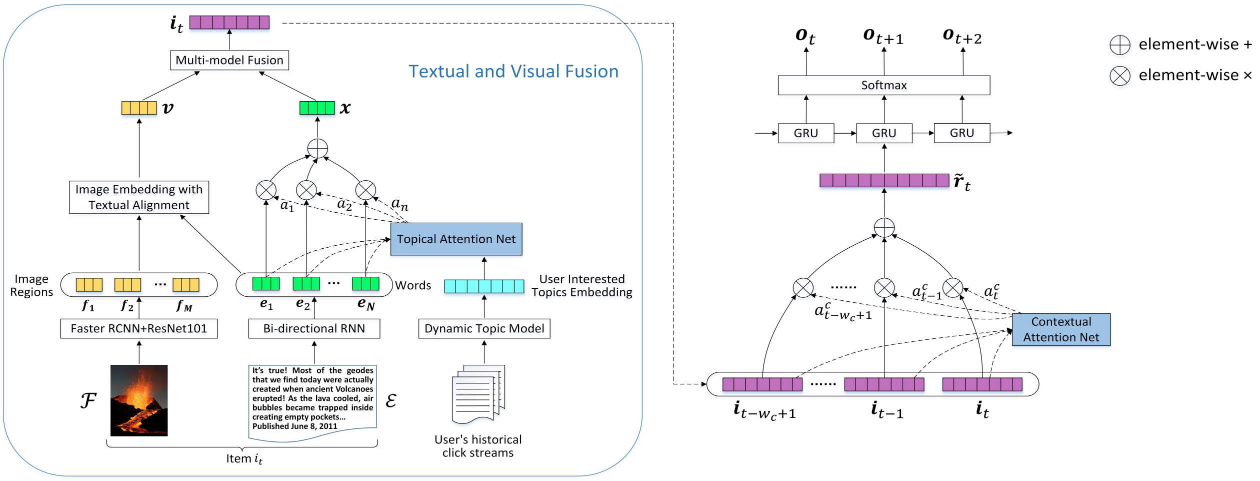


Figure 2: The proposed Ante-RNN framework for explainable recommendation. (1) In blue square, the model learns image representation $\boldsymbol{v}$ with textual alignment and text representation $\boldsymbol{x}$ with topical attention net of item at timestamp t, then the fused representation $\boldsymbol{i}_t$ can be achieved through textual and visual fusion component to denote the item embedding at timestamp t. (2) The contextual attention net takes the fused representation $\boldsymbol{i}_t$ as input to learn user's dynamic interest representation $\tilde{\boldsymbol{r}}_t$. Finally, the model outputs the probability score of the next possible item the user interacted with.

tion with ResNet-101 pre-trained by Anderson et al. (2018) to recognize the salient and obvious objects from image. Then, a fully-connected layer is added to transform each region representation into D-dimensional space. The word representations are achieved using bi-directional GRU (Schuster and Paliwal, 1997) to find the relationship between words and map language to the same dimensional semantic vector space as image regions.

Given $\mathcal{E} \in \mathbb{R}^{D\times N}$ and $\mathcal{F} \in \mathbb{R}^{D\times M}$, the image embedding with textual alignment starts with defining an affinity matrix $\boldsymbol{C} \in \mathbb{R}^{N\times M}$, whose element c_{ij} represents the similarity between the corresponding feature vector pair of $\boldsymbol{e}_i \in \mathcal{E}$ and $\boldsymbol{f}_j \in \mathcal{F}$. Specifically, $\boldsymbol{C}$ is defined as

$$\boldsymbol{C} = tanh(\mathcal{E}^T \boldsymbol{W}^b \mathcal{F}) \tag{1}$$

where $\boldsymbol{W}^b \in \mathbb{R}^{D\times D}$ denotes the correlation matrix to be learned.

Next, based on the affinity matrix, to weigh the alignment of each image region with respect to the text description, we adopt a weighted summation of all word representations denoted as

$$\boldsymbol{a}^e_j = \sum_{i=1}^{N} \alpha^e_{ij} \boldsymbol{e}_i \tag{2}$$

where $\alpha^e_{ij} = exp(c_{ij}) / \sum_{i=1}^{N} exp(c_{ij})$ denotes the weight score on how well the j-th image region and the i-th word match. After that, to determine the importance of image regions given the text description, the relevance between the j-th region and the corresponding description can be defined as

$$R(\boldsymbol{f}_j, \boldsymbol{a}^e_j) = \frac{\boldsymbol{f}^T_j \boldsymbol{a}^e_j}{||\boldsymbol{f}_j|| \cdot ||\boldsymbol{a}^e_j||} \tag{3}$$

Then, the similarity between image $\mathcal{F}$ and text description $\mathcal{E}$ can be defined as

$$S(\mathcal{E}, \mathcal{F}) = \frac{1}{M} \sum_{j=1}^{M} R(\boldsymbol{f}_j, \boldsymbol{a}^e_j) \tag{4}$$

And the representation $\boldsymbol{v}$ of image $\mathcal{F}$ can be represented as the weighted summation of all regions with respect to the alignments of text.

$$\boldsymbol{v} = \sum_{j=1}^{M} R(\boldsymbol{f}_j, \boldsymbol{a}^e_j) \cdot \boldsymbol{f}_j \tag{5}$$

In Lee et al. (2018), the authors only focus on the hardest negatives in a mini-batch when formulating the objective function. In practice, for computational efficiency, rather than summing over all the negative samples as Kiros et al. (2014), it usually considers only the hard negatives in a mini-batch of stochastic gradient descent. Thus, we define our triplet ranking loss as

$$l(\mathcal{E}, \mathcal{F}) = max[0, \alpha_1 - S(\mathcal{E}, \mathcal{F}) + S(\hat{\mathcal{E}}, \mathcal{F})] \tag{6}$$

where α_1 denotes the margin in triplet loss, $\hat{\mathcal{E}} = argmax_{t \neq \mathcal{E}} S(t, \mathcal{F})$ represents the hardest negative. Different from Lee et al. (2018), we only take into account the image-text alignment instead of both image-text and text-image alignments for that we care about how the text description can help image representations solely.

3.3. Topic-based Text Embedding

To introduce user's historical interested topics into the model learning procedure and help to learn a better representation of text description, we propose a topical attention network which incorporates topic distribution to weigh the importance and relevance of each word in the text. Specifically, we first conduct topic modelling approach on all the users' historical behaviour streams to build a shared user topic space and learn the topical distribution for each user. Users' historical behaviours are collected at a certain time interval, for instance daily, hourly and weekly. In our paper, we leverage stream LDA model introduced by Gao et al. (2016) to learn topic distributions and update the model with every user's coming streams incrementally. Therefore, the learned topic space is timely updated and can well track the recent focuses on user behaviours. After that, we aggregate all historical topic distributions of each user to derive the representation of user interested topics at the current time. Furthermore, a time decay approach (Ding and Li, 2005) is adopted to weight the different importance of the coming streams. Thus, the user's interested topics at time stamp t can be defined as:

$$\boldsymbol{\eta}_t^u = \frac{1}{N_u} \sum_{i=1}^{t} \xi_i^u \cdot e^{-\lambda|t-i|} \tag{7}$$

where ξ_i^u is the user's topic distribution at time stamp i, $|t - t_i|$ indicates the time difference between the current time and the topic time stamp i. N_u is a normalization parameter and λ is the time decay parameter.

Then, we can derive the interested topics embedding of each user u as $\boldsymbol{\eta}_t^u \in \mathbb{R}^{\psi\times 1}$ at time stamp t, where ψ is the number of topics. After that, the topical attention network outputs the text embedding $\boldsymbol{x} \in \mathbb{R}^D$ for each item i computed as a weighted summation of each word embedding e_j:

$$\boldsymbol{x} = \sum_{j=1}^{N} a_j \boldsymbol{e}_j \tag{8}$$

where D is the dimension of the word embedding, $a_j \in [0,1]$ is the attention weight of $\boldsymbol{e}_j$ and $\sum_j a_j = 1$. To obtain $a_j, j \in [1, N]$, we use the following equation to compute scores on how well the interested topics embedding $\boldsymbol{\eta}_t^u$ matches the word embedding in position j:

$$g_j = \boldsymbol{q}_a^T tanh(\boldsymbol{W}^a \boldsymbol{\eta}_t^u + \boldsymbol{U}^a \boldsymbol{e}_j) \tag{9}$$

where $\boldsymbol{W}^a \in \mathbb{R}^{D\times\psi}$, $\boldsymbol{U}^a \in \mathbb{R}^{D\times D}$ and $\boldsymbol{q}_a \in \mathbb{R}^{D\times 1}$ are the weight matrices. Finally, the topical attentive weight score a_j can be calculated with a softmax function

$$a_j = softmax(g_j) = \frac{exp(g_j)}{\sum_{j=1}^{n} exp(g_j)} \tag{10}$$

3.4. Multi-Model Fusion

In previous sections, we have described the ways to extract image and text representations, but how to model the interactions between these two features and obtain a better fusion representation is still a problem worth exploring. Therefore, in this section, we consider three different multi-modal fusion methods as shown in Figure 3 to explore the sophisticated effects.

3.4.1. Direct Fusion

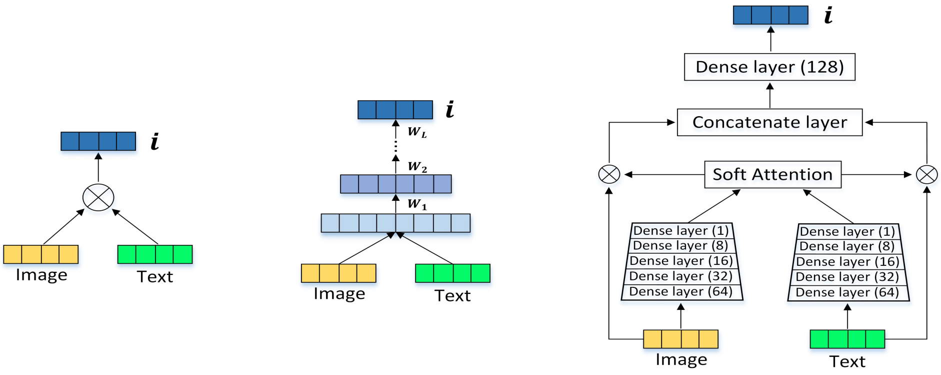
An intuitive way to do the feature fusion is to combine the learned representations of multi-modalities directly. Normally, there are three ways to fuse the learned representations, namely *concatenation*, *addition* and *element-wise product*. Here, we apply *element-wise product* which has been verified its effectiveness by Chen et al. (2018c) and reveals favored performance in our experiments.

$$\boldsymbol{i} = \boldsymbol{v} \otimes \boldsymbol{x} \tag{11}$$

where $\boldsymbol{v}$ and $\boldsymbol{x}$ denote the learned textual and visual representations and $\boldsymbol{i} \in \mathbb{R}^D$ is the output after fusion. We omit subscript t for a simpler expression.

3.4.2. Neural Fusion

Inspired by the work of Cheng et al. (2018) but differently, we first concatenate the visual and textual representations directly to keep the original modality characteristics, and then



(a) Direct Fusion

(b) Neural Fusion

(c) Attention Fusion

Figure 3: Multi-model fusion architectures.

leverage a neural network to fuse them in a complex non-linear way.

$$\boldsymbol{i} = DNN([\boldsymbol{v};\boldsymbol{x}]) \tag{12}$$

where ; represents the concatenation operation. As for $DNN(\cdot)$ model, we leverage several fully connected layers stacked together to derive the non-linear output.

$$\begin{aligned}
\boldsymbol{r}'_0 &= [\boldsymbol{v};\boldsymbol{x}] \\
\boldsymbol{r}'_1 &= \varphi(\boldsymbol{W}_1\boldsymbol{r}'_0 + \boldsymbol{b}_1) \\
\boldsymbol{r}'_2 &= \varphi(\boldsymbol{W}_2\boldsymbol{r}'_1 + \boldsymbol{b}_2) \\
&\ldots\ldots, \\
\boldsymbol{i} &= \varphi(\boldsymbol{W}_L\boldsymbol{r}'_{L-1} + \boldsymbol{b}_L)
\end{aligned}$$

where $\boldsymbol{W}_l$ and $\boldsymbol{b}_l$ denote the weight matrix and bias for the l-th fully connected layer. $\varphi(\cdot)$ denotes the activation function.

3.4.3. Attention Fusion

Same modality may have different contributions for different recommendation tasks. For instance, people show more interests on visual-related features than textual descriptions on image recommendation tasks, such as Pinterest and Instagram. While textual features might provide more useful information than other kinds of modalities in news or movie recommendations. To fully exploit the difference of multimodal nature in recommendation tasks, we apply an attention mechanism to assign different weights for multi-modalities.

Different from previous two fusions, attention fusion adopts an attention network over the extracted representations of modality-specific features, helping the recommender system to tell the different importance of the different modalities. Following the work of Gu et al. (2018b), we adopt a tower pattern network structure as the base of our attention network. The bottom layer is the widest and each successive layer has smaller number of neurons. Ultimately, the output from the last layer has the dimension of k, representing the relative importance for k different modalities. In our paper, we set $k = 2$ denoting the visual and textual modalities. Then, a softmax layer is applied to generate the weighted score for the modalities:

$$\boldsymbol{s} = softmax(TowerNet([\boldsymbol{v};\boldsymbol{x}])) \tag{13}$$

where $TowerNet(\cdot)$ represents the deep neural network with tower structure. $\boldsymbol{s} = [s_v, s_t]$ is a $k = 2$ dimensional vector representing the visual and textual attention score. Finally, a dense layer is used to learn the associations across weighted multi-modalities:

$$\boldsymbol{i} = tanh(\boldsymbol{W}_e[(1 + s_v)\boldsymbol{v}; (1 + s_t)\boldsymbol{x}] + \boldsymbol{b}_e) \tag{14}$$

where $\boldsymbol{i} \in \mathbb{R}^D$ denotes the final fused item representation. $\boldsymbol{W}_e \in \mathbb{R}^{D\times 2D}$ and $\boldsymbol{b}_e \in \mathbb{R}^D$ are parameters for the dense layer. We also keep the original modality characteristics by using $(1 + s)$.

3.5. Contextual Attention Mechanism

Given a sequence of items $\mathcal{I} = \{i_1, i_2, ..., i_t\}$ that the user u interacted with and ordered according to time, where t represents the current time stamp. Recall that $\boldsymbol{i}_t$ represents the fusion embedding of item i_t. Let $\boldsymbol{C}^t_i = [\boldsymbol{i}_{t-w_c+1}; ...; \boldsymbol{i}_t]$ be a context

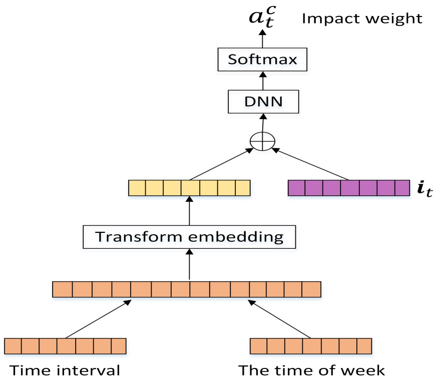


Figure 4: Diagram of contextual attention network.

matrix consisting of recent w_c inputs, where w_c is the window width of the context. To learn user's current representation considering the contextual effects, one can simply average all the representations of his/her clicked items within the contextual window:

$$\tilde{\boldsymbol{r}}_t = \frac{1}{w_c} \sum_{j=t-w_c+1}^{t} \boldsymbol{i}_j \tag{15}$$

However, user's interests are full of stochasticity and contingency, which means a user might accidentally click on wrong items or s/he is attracted by some unrelated items due to curiosity. And we argue that time plays the key role on the user's next possible behaviour. For instance, one would like to watch detective or horror movies on Friday and Saturday but might prefer comedies on other days during a week. Besides, the time interval between item i_t and item i_j, where $j < t$, also matters. Normally, events with fewer time intervals with respect to the current time have greater impact on current behaviour. Thus, apart from the representations of items within contextual window, our contextual attention mechanism also considers two other factors[1] as shown in Figure 4:

- **Time of Week $\boldsymbol{T}_w$** is the time within a week measured by hour. Specifically, we divide one week into 24 × 7 = 168 hours ordered from Monday to Sunday, and adopt a vector with 169 dimensions (the first 168 dimensions are for each hour of a week and the last one is for everything older than that) to embed the time of week $\boldsymbol{T}_w$. If a user clicked an item at such as 00:10 on Monday, then this event belongs to the first hour and the value in the first dimension of the vector will be set to 1. If an event was happened out of 168 hours, the 169th dimension will be set to 1.

- **Time Interval $\boldsymbol{\delta t}$** is the time difference between the user's historical behaviour and the current time. Similar with $\boldsymbol{T}_w$, we apply a 169-dimensional vector and the first 168 dimensions represent that the time intervals between the timestamp of previous clicked item and the current timestamp are within 0 to 168 hours. The 169-th dimension represents everything happened older than 168 hours. In this way, we can explore how time difference affects the user's next behaviour.

For each context vector $\boldsymbol{i}_j, j \in [t - w_c + 1, t]$ in $\boldsymbol{C}_i^t$, we can obtain its corresponding representation of $\boldsymbol{T}_{w,j}$ and $\boldsymbol{\delta t}_j$. To learn the two factors and item's representation $\boldsymbol{i}_j$ together, one ordinary way is the simple concatenation strategy as $[\boldsymbol{i}_j; \boldsymbol{T}_{w,j}; \boldsymbol{\delta t}_j]$. However, we argue that factor embedding and item embedding are learned differently, which means they are in different representation space. Thus, we introduce the transformed embeddings

$$\boldsymbol{i}_j^\star = g([\boldsymbol{T}_{w,j}; \boldsymbol{\delta t}_j]) \tag{16}$$

where $g(\cdot)$ is the transformation function, and can be either linear

$$g([\boldsymbol{T}_{w,j}; \boldsymbol{\delta t}_j]) = \boldsymbol{W}_f([\boldsymbol{T}_{w,j}; \boldsymbol{\delta t}_j]) \tag{17}$$

or non-linear

$$g([\boldsymbol{T}_{w,j}; \boldsymbol{\delta t}_j]) = sigmoid(\boldsymbol{W}_f([\boldsymbol{T}_{w,j}; \boldsymbol{\delta t}_j]) + \boldsymbol{b}_f) \tag{18}$$

where $\boldsymbol{W}_f \in \mathbb{R}^{D \times 338}$ ($338 = 2 \times 169$) is the trainable transformation matrix and $\boldsymbol{b}_f \in \mathbb{R}^{D \times 1}$ is the trainable bias. Since the transformation is continuous, it can map factor embeddings to item space while preserving their original relationship. We therefore can concatenate these two embeddings as $\tilde{\boldsymbol{i}}_j = [\boldsymbol{i}_j^\star; \boldsymbol{i}_j]$.

[1] It is worth noting that other kinds of factors such as location can also be considered and according to the same transformation mechanism but it is beyond the scope of this paper.

After that, we perform the following attention mechanism:

$$a_j^c = softmax(\mathcal{G}(\tilde{\boldsymbol{i}}_j)) = \frac{exp(\mathcal{G}(\tilde{\boldsymbol{i}}_j))}{\sum_{t-w_c+1\leq k\leq t} exp(\mathcal{G}(\tilde{\boldsymbol{i}}_k))} \tag{19}$$

where $\mathcal{G}$ is a deep neural network regarded as attention network and $softmax(\cdot)$ is the softmax function to calculate the normalized impact weight. The attention network $\mathcal{G}$ receives concatenation embedding as input and outputs the impact weight. Finally the embedding of user's current representation can be calculated as weighted summation of all item embeddings within the contextual window:

$$\tilde{\boldsymbol{r}}_t = \sum_{t-w_c+1\leq k\leq t} a_k^c \boldsymbol{i}_k \tag{20}$$

We will demonstrate the efficacy of the attention network in the experiment section.

3.6. Ante-RNN Model

Long Short-Term Memory (LSTM) is a special form of RNN, widely used to model sequence data. LSTM uses input gate, forget gate and output gate vectors at each position to control the passing of information along the sequence and thus improves the modelling of long-range dependencies. Gated Recurrent Unit (GRU) is the simplified version of LSTM networks but still maintains all their properties (Cho et al., 2014). In GRU unit, the activation h_t at time t is a linear interpolation between the previous activation h_{t-1} and the candidate activation $\tilde{h}_t$. After we get the output vector $\tilde{\boldsymbol{r}}_t$ from contextual attention layer as the input to the GRU layer, the following intermediate calculations can be achieved recursively during model learning procedure:

$$\begin{aligned}
\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z\tilde{\boldsymbol{r}}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1}) \\
\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r\tilde{\boldsymbol{r}}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1}) \\
\tilde{\boldsymbol{h}}_t &= tanh(\boldsymbol{W}_c\tilde{\boldsymbol{r}}_t + \boldsymbol{U}_c(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1})) \\
\boldsymbol{h}_t &= (1 - \boldsymbol{z}_t)\boldsymbol{h}_{t-1} + \boldsymbol{z}_t\tilde{h}_t
\end{aligned} \tag{21}$$

where update gate $\boldsymbol{z}_t$ decides how much the unit updates its activation or content. $\boldsymbol{r}_t$ is a set of reset gate to control the flow of information, and $\odot$ is an element-wise multiplication. $\sigma(\cdot)$ and $tanh(\cdot)$ are the element-wise logistic function and hyperbolic tangent function used to do non-linear projection. The length of the output vector $\boldsymbol{o}_t$ from GRU layer is the number of all candidate items, and a softmax layer is added after GRU layer to output the probability distributions of all candidate items.

Illuminated by the recent successes of probabilistic sequential translation model (Pan et al., 2016), given a set of user's interacted items $\mathcal{I}_u = \{\boldsymbol{i}_1, \boldsymbol{i}_2, ..., \boldsymbol{i}_t\}$ and current user's interested topics $\boldsymbol{\eta}_t^u$, we formulate our recommendation problem as a coherence loss, where the log probability of the recommendation is given by the sum of log probabilities over the clicked items:

$$l(\boldsymbol{\eta}_t^u, \mathcal{I}_u) = -logPr(\mathcal{I}_u|\boldsymbol{\eta}_t^u) = \sum_{t=1}^{N_u} -logP(\boldsymbol{i}_t|\boldsymbol{\eta}_t^u, \boldsymbol{i}_1, \boldsymbol{i}_2, ..., \boldsymbol{i}_{t-1}; \Theta) \tag{22}$$

where $\{\boldsymbol{i}_1, \boldsymbol{i}_2, ..., \boldsymbol{i}_{N_u}\}$ is the sequentially predicted items. Here, $\boldsymbol{i}$ is corresponding to the fused image and textual embedding. By performing our contextual attention mechanism, for each time stamp t, we can get the user interest embedding $\tilde{\boldsymbol{r}}_t \in \mathbb{R}^D$ as the GRU input, as shown in Figure 2. Θ is the set of parameters of our framework, including contextual attention and GRU layer, the parameters in image embedding network, topical attention network and multi-model fusion component. By minimizing the above loss, the user interest evolvement can be described dynamically, making the recommendation more coherent and reasonable. Then the above probabilities can be achieved through softmax classification function demonstrated below:

$$P(\boldsymbol{i}_t = p|\boldsymbol{\eta}_t^u, \boldsymbol{i}_1, \boldsymbol{i}_2, ..., \boldsymbol{i}_{t-1}; \Theta) = \frac{exp(\boldsymbol{W}_s^{(p)}\boldsymbol{h}_t)}{\sum_{j=1}^{|\mathcal{I}|} exp(\boldsymbol{W}_s^{(j)}\boldsymbol{h}_t)} \tag{23}$$

where $|\mathcal{I}|$ is the number of candidate items, $\boldsymbol{W}_s$ is the parameter matrix of the softmax layer in our model.

Finally, we can obtain the objective function as:

$$\mathcal{L} = \sum_{u\in\mathcal{U}} l(\boldsymbol{\eta}_t^u, \mathcal{I}_u) + \lambda_1 \sum_{\mathcal{E},\mathcal{F}} l(\mathcal{E}, \mathcal{F}) + \lambda_2||\Theta||_2^2 \tag{24}$$

where λ_1 is the trade-off parameters for these objectives. $\lambda_2 \geq 0$ is the coefficient of the weight decay term. Then, Ante-RNN can be learned by the stochastic gradient descent and BPTT. The parameters are automatically updated by Theano (Bergstra et al., 2010). By optimizing the above overall loss function in a unified framework, our proposed method achieves personalized dynamic recommendation with considering image-textual alignment, user's interested topics, multi-model fusion and contextual influence jointly.

4. Experiments

In this section, we conduct our experiments on two real-world datasets. First, we introduce the datasets, evaluation metrics and baseline methods as well as parameter settings. Then we make comparison between Ante-RNN model and the baselines. After that, the recommendation efficiency and the effectiveness of the hybrid attention mechanism proposed in this paper will be tested, followed by the analysis on users with different sparsity level and various parameters. Finally, we illustrate the recommendation explainability.

4.1. Datasets

Experiments are conducted on two large scale datasets, namely Movielens [2] and Pinterest [3]. The basic statistics of them are listed in Table 3. For both datasets, we sort all user-item interaction pairs in the ascending interaction time order. The first 80% sequential histories are selected as training set and the rest 20% as test set. Besides, we randomly hold-out 10% interaction history of each user from training set as validation sets. To measure the statistical significance of Ante-RNN over the baselines, we repeat the splitting process five times (i.e., generating five pairs of training and validation sets). Averaged results are reported in the following subsection.

Table 3: Main properties of the experimental datasets.

Dataset	*#Users*	*#Items*	*#Interactions*	*#Avg. seq. len.*	*#Sparsity*
MovieLens	283,228	58,098	27,753,444	115	99.83%
Pinterest	50,000	14,965	1,091,733	23	99.85%

1. MovieLens dataset contains 27,753,444 ratings from 283,228 users on 58,098 movies from January 09, 1995 to September 26, 2018. In order to mimic implicit data, we binarized all ratings independent of their values, considering them as positive feedback as it has been done by Rendle et al. (2009). Using the timestamps provided, we thus got an ordered sequence of consumption events for each user. The dataset contains only sequences with a minimum length of 20. The average sequence length is 115. We aimed at predicting the next movie to watch. In order to obtain the textual information and poster image corresponding to each movie, we downloaded descriptions and images according to *tmdbID* property provided in *links.csv* file through TMDb API[4].

2. Pinterest is one of the largest social curation networks. This dataset with implicit feedback is constructed by Geng et al. (2015) for evaluating image recommendation. Due to the large volume and high sparsity of this dataset, for instance, over 20% of users have only one pin, we filter the dataset by retaining the top 15,000 popular images and sampling 50,000 users who have interactions on these images. This results in a subset of data that contains 50,000 users, 14,965 images and 1,091,733 interactions. Each interaction denotes whether the user has pinned the image to his/her own board. Since there is no description information on images, we also collect corresponding descriptions by using Pinterest API[5].

A sample of the dataset can be accessible through the link[6], and the full version of our dataset is available on request.

4.2. Evaluation Metrics

Based on temporally ordered lists of pinned/rated items, our objective is to correctly predict the next item a target user will likely pin/rate. The ground truth at a particular time step is therefore represented by a single user-item tuple. To present the user with adequate recommendations, the target item should

[2] http://files.grouplens.org/datasets/movielens/ml-latest-README.html

[3] https://sites.google.com/site/xueatalphabeta/academic-projects.

[4] https://www.themoviedb.org/documentation/api

[5] https://developers.pinterest.com/docs/api/

[6] `https://www.dropbox.com/sh/hinouvmaj7lginn/AABpgBifZLQBYrHLHaVzzSUQa?dl=0`

be among the top few recommended items. Since we are interested in measuring top-K recommendation instead of rating prediction, we measure the quality by looking at the *Recall@K* and *NDCG@K*, which are widely used for evaluating top-K recommender systems.

- *Recall@K* is defined as the fraction of cases where the item actually consumed in the next event is among the top K items recommended (Powers, 2011).
- *NDCG@K* (Normalized Discounted Cumulative Gain) is adopted to evaluate ranking performance by taking the positions of the correct items into consideration (Järvelin and Kekäläinen, 2000), and thus to assess if the items that a user has actually consumed are ranked in higher positions in the recommendation list.

We set $K = 20$, as it appears desirable from a user's perspective to expect the target among the first 20 items (Hidasi et al., 2015).

4.3. Baselines

To validate the effectiveness of Ante-RNN, we compared our model with the following methods. Note that all model-based Collaborative Filtering approaches are learned by optimizing the same pairwise ranking loss of Bayesian Personalized Ranking (BPR) for a fair comparison. BPR will be introduced in detail below.

- **BPR**[7]: This method optimizes the latent factor model with a pairwise ranking loss, which is tailored to learn from implicit feedback. It is a highly competitive and popular baseline for item recommendation (Rendle et al., 2009). We adopt matrix factorization as the prediction component for BPR.
- **VBPR**[8]: The Visual Bayesian Personalized Ranking (VBPR) model is a state-of-the-art method for recommendation leveraging item visual images (He and McAuley, 2016).
- **CTR**[9]: Collaborative Topic Regression (CTR) learns interpretable latent structure from user generated contents so that probabilistic topic modelling can be integrated into collaborative filtering (Wang and Blei, 2011).
- **GRU**[10]: It is the state-of-the-art sequential recommendation method, and an extension of RNN for capturing the long-term dependency (Yu et al., 2016). GRU is also the basic of our Ante-RNN model.
- **IARN**[11]: Interacting Attention-gated Recurrent Network (IARN) model proposed by Pei et al. (2017) integrates an attention mechanism into BRNN when modelling both user and item representations for the sequential recommendation. Then the inner product of user and item representations is performed to predict user ratings.
- **MLAM**[12]: The Multi-level Attraction Model (MLAM) is a state-of-the-art interpreterable recommendation algorithm, which leverages attention-based multi-level contextual information for Top-K recommendation (Hu et al., 2018) and meanwhile provides explanations. In our situation, we apply image features instead of the cast level module and then build attractions over them.
- **MV-RNN**[13]: Multi-View Recurrent Neural Network (MV-RNN) proposed by Cui et al. (2018) is a newly proposed algorithm especially for sequential recommendations. Similarly, it incorporates visual and textual information to deal with cold start issue and meanwhile applies a recurrent structure to dynamically capture the users' interests. Differently, they do not consider time factors between user's historical interactions and they use a denoising autoencoder for multi-modality fusion.

Besides, we also adopt two variations of our Ante-RNN model, namely **t-Ante-RNN** and **v-Ante-RNN**. In former model,

[7] https://github.com/gamboviol/bpr.

[8] https://sites.google.com/a/eng.ucsd.edu/ruining-he/.

[9] https://github.com/blei-lab/ctr.

[10] https://github.com/LaceyChen17/DREAM.

[11] https://github.com/wenjiepei/IARN.

[12] https://github.com/rainmilk/ijcai18mlma.

[13] https://github.com/cuiqiang1990/MV-RNN

we only keep text description of item as input and remove all modules that are related to image processing to perform recommendations, whereas the latter one only leverages images as model inputs and modules with respect to text processing are excluded when generating Top-N rank list. Other two variations of Ante-RNN are **Ante-RNN-D**, **Ante-RNN-N** represent Ante-RNN with direct fusion and neural fusion respectively, while we use Ante-RNN to represent Ante-RNN with attention fusion for it achieves the best performance of all fusion methods.

4.4. Parameter Settings

For image embedding of Ante-RNN model, we use Faster R-CNN in conjunction with ResNet-101 pre-trained by Anderson et al. (2018) to extract Region Of Interests (ROIs) for each image. The Faster R-CNN implementation uses an intersection over union (IoU) threshold of 0.7 for region proposal suppression, and 0.3 for object class suppression. The class detection confidence threshold is set as 0.2 to select salient image regions, and top 36 ROIs with highest confidence scores are selected. We extracted features after average pooling, resulting in the final representation of 2048 dimensions. The embedding dimension D is set to 128. Topic numbers ψ is set to 70 and 100 for MovieLens and Pinterest datasets respectively. For time decay rate λ, we set it to 0.2 for MovieLens dataset, but a relatively slow decay $\lambda = 0.1$ for Pinterest dataset. In the model training phase, the trade-off parameter λ_1 is set to 0.2 by grid-search over $\{0.2, 0.4, 0.6, 0.8\}$. The coefficient λ_2 of weight decay term is set to 0.0001. The contrastive margin α_1 is set to 0.3. Learning rate is set to 0.001. The window sizes w_c are set as 5 and 3 for MovieLens and Pinterest dataset respectively.

The hyper-parameters of each baseline are tuned with the validation set during training phase. Specifically, the dimension of latent factors (or embedding size) is set to 128 for baselines. The regularization coefficient is set to 10 that works best for BPR and VBPR. We set α of MLAM to 1, 4 and 2 for image, word and sentence level attention model. Optimization for baselines terminate until convergence or 150 learning epochs. Other parameters are set the same with our model if not specified.

4.5. Performance Evaluation

The performance of Ante-RNN and the baselines are reported in terms of Recall@K and NDCG@K on two kinds of datasets in Table 4. K ranges over $\{5, 10, 15, 20, 25\}$. From the results, we can see that: 1) The performance of BPR fails to surpass the rest baseline models since that the latter ones integrate either visual or text features into their modelling process. This observation verifies that side information, e.g. image or text, is complementary to ratings/ implicit feedbacks and thus can help to improve recommendation performance in real-world applications. Furthermore, by incorporating both visual and textual information, MLAM, MV-RNN and our Ante-RNN models obtain the best performance among all comparison methods. 2) It is worth noting that different side information takes on different importance for different datasets. For instance, t-Ante-RNN achieves better performance than v-Ante-RNN on MovieLens while performs worse than v-Ante-RNN on Pinterest. The reason may be that for movie recommendations, the users pay more attention to the plot and descriptions on movies instead of posters, while users on Pinterest focus more on images than other side information. 3) For the baselines, neural recommendation algorithms, namely MLAM, IARN, GRU, MV-RNN, Ante-RNN and its variations, greatly perform better than the other baselines for that they can either better learn the latent features of items or better model user's dynamic interests over time from sequential inputs. Among these, the performance of MV-RNN is better than MLAM and IARN which verifies the importance of both capturing user's sequential patterns and integrating multiple side information. 4) Our Ante-RNN model outperforms over the baselines on all datasets and evaluation measures by combining visual and textual information into representation learning process. Furthermore, the performance of Ante-RNN is better than MV-RNN because the hybrid attention mechanism also helps to model user's long and short-term dynamic preferences. On MovieLens, it outperforms the best baseline MV-RNN by 4.6% on Recall@20 and 4.1% on

Table 4: Performance comparison (Mean ± Standard Deviation) w.r.t. Recall@K and NDCG@K (K=5, 10, 15, 20, 25) on two datasets (MovieLens and Pinterest). "*" indicates that the improvements of our model over the best baseline are statistically significant for p-value < 0.01 with paired t-test.

MovieLens	**Recall@5**	**Recall@10**	**Recall@15**	**Recall@20**	**Recall@25**	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**NDCG@25**
BPR	0.128 ± 0.021	0.154 ± 0.012	0.172 ± 0.011	0.199 ± 0.016	0.215 ± 0.014	0.083 ± 0.009	0.096 ± 0.008	0.102 ± 0.011	0.114 ± 0.021	0.127 ± 0.012
CTR	0.189 ± 0.024	0.206 ± 0.013	0.223 ± 0.019	0.237 ± 0.021	0.251 ± 0.018	0.096 ± 0.011	0.103 ± 0.009	0.115 ± 0.015	0.128 ± 0.006	0.142 ± 0.013
VBPR	0.211 ± 0.018	0.226 ± 0.023	0.240 ± 0.021	0.256 ± 0.015	0.274 ± 0.019	0.102 ± 0.013	0.116 ± 0.012	0.129 ± 0.019	0.141 ± 0.011	0.153± 0.016
GRU	0.261 ± 0.031	0.274 ± 0.026	0.298 ± 0.019	0.311 ± 0.021	0.326 ± 0.024	0.133 ± 0.011	0.142 ± 0.016	0.157 ± 0.009	0.173 ± 0.018	0.185 ± 0.014
IARN	0.282 ± 0.016	0.297 ± 0.022	0.316 ± 0.034	0.331 ± 0.023	0.343 ± 0.021	0.157 ± 0.015	0.163 ± 0.011	0.179 ± 0.019	0.192 ± 0.008	0.208 ± 0.017
MLAM	0.309 ± 0.033	0.321 ± 0.026	0.342 ± 0.029	0.358 ± 0.018	0.371 ± 0.013	0.174 ± 0.011	0.186 ± 0.014	0.203 ± 0.012	0.217 ± 0.017	0.232 ± 0.015
MV-RNN	0.326 ± 0.023	0.338 ± 0.018	0.361 ± 0.019	0.375 ± 0.022	0.389 ± 0.025	0.191 ± 0.014	0.204 ± 0.011	0.225 ± 0.009	0.231 ± 0.015	0.253 ± 0.013
v-Ante-RNN	0.273 ± 0.026	0.291 ± 0.021	0.309 ± 0.023	0.326 ± 0.016	0.337 ± 0.020	0.146 ± 0.013	0.157 ± 0.011	0.162 ± 0.016	0.184 ± 0.012	0.203 ± 0.014
t-Ante-RNN	0.301 ± 0.018	0.315 ± 0.016	0.329 ± 0.019	0.347 ± 0.021	0.364 ± 0.022	0.162 ± 0.011	0.174 ± 0.015	0.193 ± 0.009	0.209 ± 0.014	0.226 ± 0.011
Ante-RNN-D	0.342 ± 0.013	0.361 ± 0.017	0.384 ± 0.008	0.393 ± 0.016	0.407 ± 0.014	0.208 ± 0.014	0.221 ± 0.012	0.239 ± 0.017	0.252 ± 0.013	0.275 ± 0.016
Ante-RNN-N	0.359 ± 0.008	0.385 ± 0.016	0.401 ± 0.021	0.416 ± 0.014	0.429 ± 0.017	0.220 ± 0.013	0.238 ± 0.018	0.251 ± 0.007	0.265 ± 0.016	0.286 ± 0.014
Ante-RNN	**0.365 ± 0.006***	**0.389± 0.013***	**0.408± 0.016***	**0.421± 0.015***	**0.436± 0.011***	**0.227± 0.012***	**0.246± 0.007***	**0.263± 0.016***	**0.272± 0.021***	**0.298± 0.015***

Pinterest	**Recall@5**	**Recall@10**	**Recall@15**	**Recall@20**	**Recall@25**	**NDCG@5**	**NDCG@10**	**NDCG@15**	**NDCG@20**	**NDCG@25**
BPR	0.056 ± 0.018	0.073 ± 0.014	0.085 ± 0.011	0.094 ± 0.016	0.107 ± 0.023	0.032 ± 0.013	0.039 ± 0.011	0.047 ± 0.009	0.052 ± 0.014	0.058 ± 0.018
CTR	0.071 ± 0.011	0.083 ± 0.014	0.092 ± 0.017	0.106 ± 0.012	0.124 ± 0.015	0.039 ± 0.012	0.048 ± 0.016	0.053 ± 0.015	0.062 ± 0.019	0.067 ± 0.021
VBPR	0.078 ± 0.013	0.092 ± 0.024	0.103 ± 0.018	0.114 ± 0.012	0.135 ± 0.016	0.042 ± 0.008	0.053 ± 0.014	0.059 ± 0.017	0.064 ± 0.011	0.071 ± 0.019
GRU	0.109 ± 0.024	0.120 ± 0.013	0.131 ± 0.011	0.142 ± 0.016	0.153 ± 0.021	0.058 ± 0.012	0.061 ± 0.011	0.066 ± 0.016	0.071 ± 0.014	0.076 ± 0.012
IARN	0.113 ± 0.016	0.126 ± 0.012	0.139 ± 0.020	0.152 ± 0.017	0.164 ± 0.014	0.061 ± 0.008	0.067 ± 0.012	0.072 ± 0.011	0.078 ± 0.017	0.085 ± 0.014
MLAM	0.151 ± 0.019	0.173 ± 0.021	0.186 ± 0.017	0.201 ± 0.013	0.218 ± 0.014	0.079 ± 0.010	0.085 ± 0.014	0.093 ± 0.013	0.102 ± 0.016	0.114 ± 0.020
MV-RNN	0.175 ± 0.017	0.188 ± 0.015	0.207 ± 0.008	0.219 ± 0.012	0.237 ± 0.011	0.093 ± 0.016	0.101 ± 0.014	0.108 ± 0.019	0.119 ± 0.021	0.130 ± 0.017
t-Ante-RNN	0.136 ± 0.017	0.142 ± 0.023	0.157 ± 0.018	0.169 ± 0.015	0.183 ± 0.013	0.068 ± 0.011	0.074 ± 0.016	0.079 ± 0.016	0.091 ± 0.013	0.098 ± 0.011
v-Ante-RNN	0.141 ± 0.022	0.154 ± 0.015	0.162 ± 0.017	0.177 ± 0.016	0.197 ± 0.019	0.071 ± 0.010	0.076 ± 0.007	0.083 ± 0.015	0.095 ± 0.013	0.103 ± 0.016
Ante-RNN-D	0.202 ± 0.018	0.216 ± 0.022	0.229 ± 0.015	0.245 ± 0.011	0.258 ± 0.014	0.109 ± 0.012	0.122 ± 0.015	0.126 ± 0.011	0.137 ± 0.019	0.154 ± 0.013
Ante-RNN-N	0.218 ± 0.012	0.231 ± 0.016	0.241 ± 0.016	0.257 ± 0.013	0.269 ± 0.018	0.124 ± 0.008	0.136 ± 0.014	0.141 ± 0.012	0.149 ± 0.014	0.166 ± 0.012
Ante-RNN	**0.223 ± 0.011***	**0.238± 0.014***	**0.252± 0.008***	**0.264± 0.015***	**0.276± 0.013***	**0.129± 0.010***	**0.145± 0.008***	**0.148± 0.016***	**0.162± 0.014***	**0.181± 0.014***

NDCG@20, and much higher than the other baseline models.

Besides, we also analyze Ante-RNN with three fusion methods. From the table, we can observe that Ante-RNN-N always beats the Ante-RNN-D. It is because the non-linear transformation boosts the interactions among multi-modalities which leads to a better fusion. The best performance appears with attention fusion but the advantage is not prominent. It indicates that the attention mechanism is more likely able to better capture the different importance of multiple input features.

4.6. Recommendation Efficiency

In addition to the advantage of recommendation accuracy, we have also evaluated the efficiency of Ante-RNN on both datasets. Table 5 shows the runtime comparison with GRU, IARN and MV-RNN. Other baselines are not listed here as the implementation cannot leverage the computation power of GPU. Experiments were conducted on a machine with a NVIDIA TITAN X Pascal GPU. From this results, we observe that Ante-RNN is comparable with other state-of-the-art approaches not utilizing the image information. Moreover, due to the efficient sampling strategy for image-text alignment, our method converges faster than MV-RNN which integrates image and text features by using autoencoder.

During prediction process, given user clicking item i_t at time stamp t, the image embedding with textual alignment v and word representations $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_N\}$ of its corresponding description can be achieved beforehand. For user u, the user interested

Table 5: Runtime comparison (seconds) for training model on both datasets.

Dataset	*GRU*	*IARN*	*Ante-RNN*	*MV-RNN*
MovieLens	3725.03	4306.82	4913.67	12416.51
Pinterest	918.34	1150.96	1431.29	3504.02

topics embedding $\boldsymbol{\eta}_t^u$ can also be derived separately according to a certain time interval, hour for instance according to equation 7. Therefore, the actual online prediction can be accelerated by only performing basic matrix operations with GPU.

4.7. Effect of Attention Mechanism

To get a better understanding of our Ante-RNN model, we further evaluate the key component - topical (T) and contextual (C) attention mechanisms. In order to prove the importance of time factors, we also evaluate two variations of contextual attention mechanism, C-$\boldsymbol{T_w}$ and C-$\boldsymbol{\delta t}$, by removing the time of week or the time interval parameter from contextual attention network. Table 6 shows the effect of our basic Ante-RNN model with or without attention mechanism(s) for $K = 20$. Note that: when we consider neither topical attention or contextual attention mechanism, it means we only adopt image embedding fused with textual embedding as GRU input for model learning, and the text embedding is the average of word representation in the text. From the table, we can observe that:

(1) When both topical and contextual attention mechanisms are applied, the recommendation performance is improved compared with the other combinations. The good performance of attention mechanism shows that the characteristics of user's long-term and short-term interests are reflected at both levels.

(2) The contextual attention mechanism contributes more for our model on two datasets as compared to topic-based attention mechanism since the performance of our model deteriorates more without contextual attention component. This may be due to the fact that the contextual attention method can strengthen the user's short-term interest modelling which GRU may lack, and capture the user's main focus during a limited time period, while the topic-based attention mechanism can assist GRU to model user's long-term interest pattern in a better way, which also leads to the improvement of recommendation performance compared with the model without topic-based attention. Furthermore, two kinds of time factors integrated in contextual attention method further strengthen the discriminating ability of user's short-term focus.

Table 6: Effect of topical (T) and contextual (C) attention mechanisms as well as their variations w.r.t. Recall@20 and NDCG@20. "*" indicates the statistical significance for p-value < 0.01.

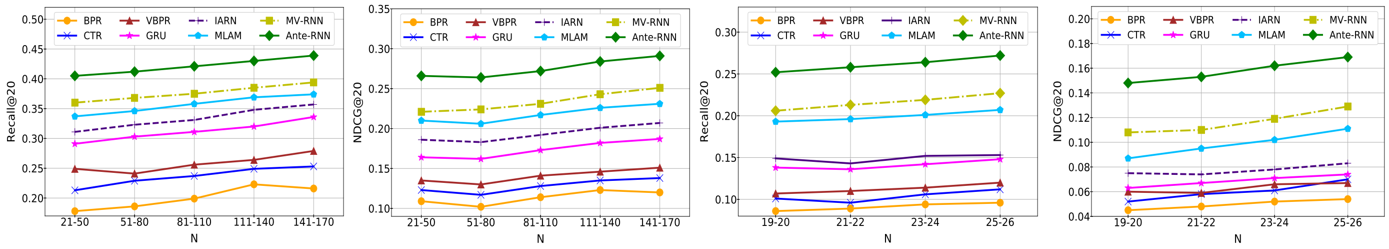
Model	Attention Type	**MovieLens**		**Pinterest**	
		Recall@20	*NDCG@20*	*Recall@20*	*NDCG@20*
Ante-RNN	None	0.364	0.228	0.205	0.112
	T	0.385	0.246	0.227	0.133
	C-$\boldsymbol{T_w}$	0.398	0.253	0.246	0.147
	C-$\boldsymbol{\delta t}$	0.387	0.249	0.232	0.136
	C	0.406	0.259	0.253	0.151
	T+C-$\boldsymbol{T_w}$	0.414	0.268	0.258	0.159
	T+C-$\boldsymbol{\delta t}$	0.409	0.261	0.251	0.148
	T+C	**0.421***	**0.272***	**0.264***	**0.162***

(3) When time interval is removed from contextual attention network, the recommendation performance deteriorates more than the contextual attention without time of week. For example, comparing to the attention network C, the performance degradation of C-$\boldsymbol{\delta t}$ is 1.9% and 2.1% on Recall@20 in MovieLens and Pinterest datasets respectively, while the performance degradation of C-$\boldsymbol{T_w}$ is 0.8% and 0.7% correspondingly. It demonstrates that time interval is more important to capture the user's short-term interest compared with time of week.

4.8. Analysis on Users with Different Sparsity Levels

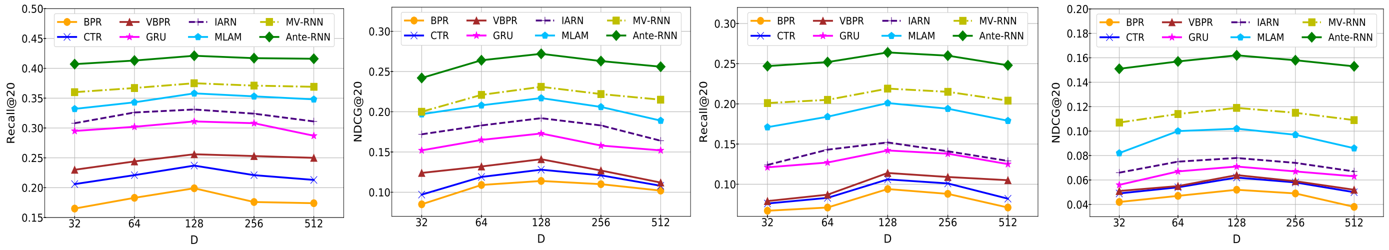
In this section, we study the impact of different sequence lengths on the recommendation performance. Note that we do not retrain our model with different sets of users, instead we divide the test set into different groups by the number of items per user. The results are shown in Figure 5, and we have the following observations:

(1) As sequence length increases, the performance of all methods generally improves, indicating that sufficient temporal context could ensure models that capture user's interest patterns in a better way. This also explains why the overall performance



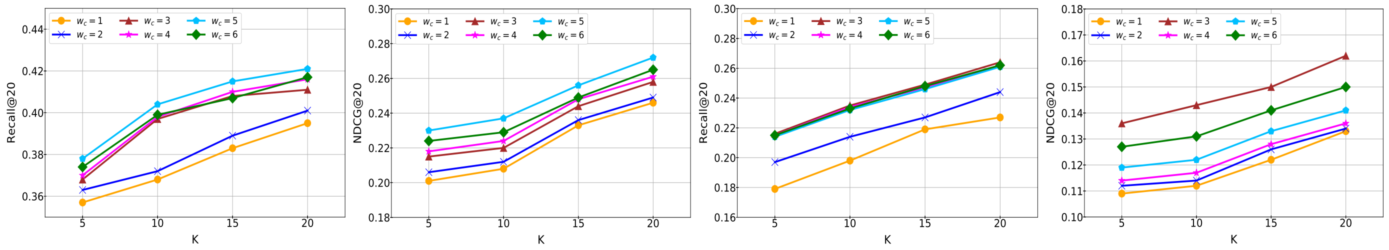
(a) MovieLens-Recall@20 (b) MovieLens-NDCG@20 (c) Pinterest-Recall@20 (d) Pinterest-NDCG@20

Figure 5: Performance of Recall@20 and NDCG@20 w.r.t. the number of items per user on two datasets.



(a) MovieLens-Recall@20 (b) MovieLens-NDCG@20 (c) Pinterest-Recall@20 (d) Pinterest-NDCG@20

Figure 6: Performance of Recall@20 and NDCG@20 w.r.t. the embedding size $D \in \{32, 64, 128, 256, 512\}$ on two datasets.



(a) MovieLens-Recall@20 (b) MovieLens-NDCG@20 (c) Pinterest-Recall@20 (d) Pinterest-NDCG@20

Figure 7: Performance of Recall@20 and NDCG@20 w.r.t. the window size $w_c \in \{1, 2, 3, 4, 5, 6\}$ on two datasets.

on MovieLens dataset is better than that on Pinterest.

(2) Overall, Ante-RNN achieves the best performance across all different configurations of all the datasets, especially, when the sequence length gets larger. On average, the relative improvements w.r.t. the second best method are 4.5% with training records length of 21-50 and 4.3% with training records length of 25-26 on Recall@20 in MovieLens and Pinterest datasets respectively. This implies the remarkable advantage of Ante-RNN in dealing with long sequences. Besides, we also find that when the number of items per user is relatively small, Ante-RNN still keeps the advantages in performance, which indicates that the hybrid attention mechanism and visual information integration could improve the recommendation quality when there is insufficient training data for each user.

4.9. *Parameter Analysis*

In this section, we analyse the influence of the embedding size D and window length w_c in the contextual attention mechanism to the performance of our Ante-RNN model.

4.9.1. *Analysis of Embedding Size D*

The empirical results displayed in Figure 6 indicate the substantial influence of embedding size upon Ante-RNN and other baselines. During experiments, we range D in {32, 64, 128, 256, 512} and fix other hyper-parameters to plot the corresponding results for $K = 20$ with respect to Recall@K and NDCG@K on two datasets. Similar trends can be found on all models of both datasets. On one hand, the recommendation performance improves along with the increasing of dimensionality, which means that the representations hold more and informative resources extracted from users and items. On the other hand, the performance of recommendations will drop when the dimensionality continues raising, which demonstrates that the models may suffer from over-fitting problem. It is worth noting that the performance of our Ante-RNN model only slightly deteriorates compared with other methods for that our model holds a higher stability for the changes of dimensionality.

4.9.2. Analysis of Window Size w_c

In this part, we investigate the best window size for Ante-RNN. The window size w_c ranges from 1 to 6 with other hyper-parameter fixed. When w_c is set as 1, it can be considered as the special case without contextual attention mechanism in our experimental cases. Figure 7 shows the performance results for $K = 20$ with respect to Recall@K and NDCG@K on MovieLens and Pinterest datasets. For both datasets, slightly difference can be observed on Recall@20 when $w_c > 2$ while there is an obvious difference on NDCG@20. We can also observe that the best window size can be chosen as $w_c = 3$ and $w_c = 5$ on Pinterest and MovieLens respectively. The superior window size on MovieLens is larger than that of Pinterest, which may be because the average sequence length on MovieLens is longer.

4.10. Recommendation Explainability

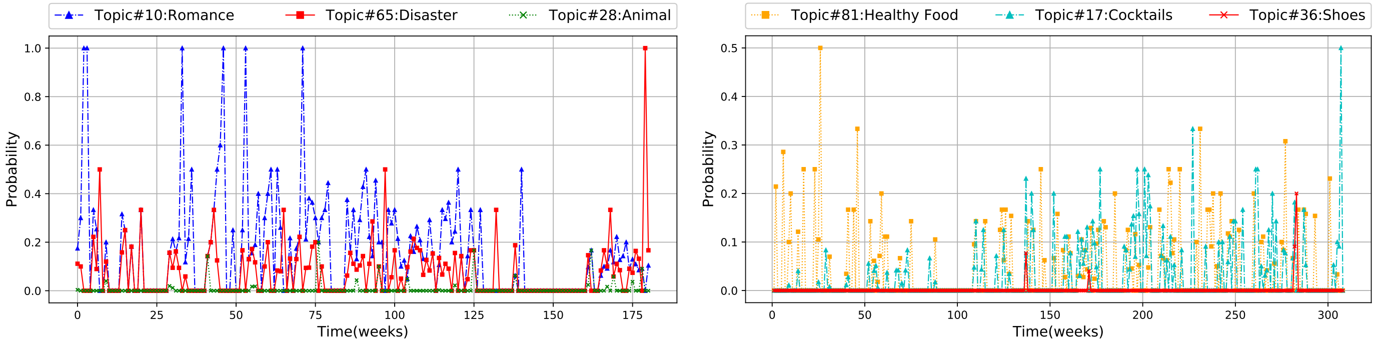
In this section, we evaluate the explanations generated by Ante-RNN from both qualitative and quantitative perspectives based on the Movielens and Pinterest datasets.

4.10.1. Qualitative Evaluation

To provide better intuitions for the generated multi-model explanations of our recommendation results and to provide a better understanding of our hybrid attention mechanism, we present and analyze two examples learned by the model in a qualitative manner. We also compare our method with MLAM, a state-of-the-art explainable recommendation algorithm on two datasets. The examples are shown in Figure 9. In particular, we show one user for each dataset with their topic historical information on the left side. The user's recent pinned/rated four items are displayed according to time order as well as the topics that they belong to and their corresponding top-4 topic words extracted from item descriptions. The number on top of each image represents the weight calculated from contextual attention layer and higher value means that the item is more important in next recommendation task. When the model has predicted the next possible item i_{t+1}, the text description of i_{t+1} will be compared with user's interested topics and related topic words are tagged with red in our examples. Then, the highlighted regions in red square of item images are determined by performing Eq.(1) and Eq.(2).

The changes of user's interested topic distribution across different weeks of two datasets are shown in Figure 8. Here we only select 3 representative topics appeared in users' recent historical records to illustrate the dynamic nature of users' interests, or else there will be too many lines tangled in one figure. We can see that, for example, on MovieLens, user 1 shows his/her long-term interests on topic #10 about "Romance" movies (blue line in Figure 8(a)) which frequently occurred in his/her historical records. However, the user's current interests shift to topic #65 "Disaster" movies (red line in Figure 8(a)) and topic #28 "Animal" movies (green line in Figure 8(a)). Our model can capture user's real-time interests through the dynamic contextual attention mechanism and recommend "Disaster" related movie. Some of the highlighted topic words i.e. *storm* and *seas*, can also be found in user's visited items. However, though MLAM can also provide explanations on image and its description separately, they cannot align highlighted image region together with its significant words. Besides, MLAM thinks the major interest of user 1 is "Romance" movies (topic #10) and thus recommends *Remember Me* instead, which verifies that our model can capture users' dynamic preferences, whereas MLAM can only model static users' interests.

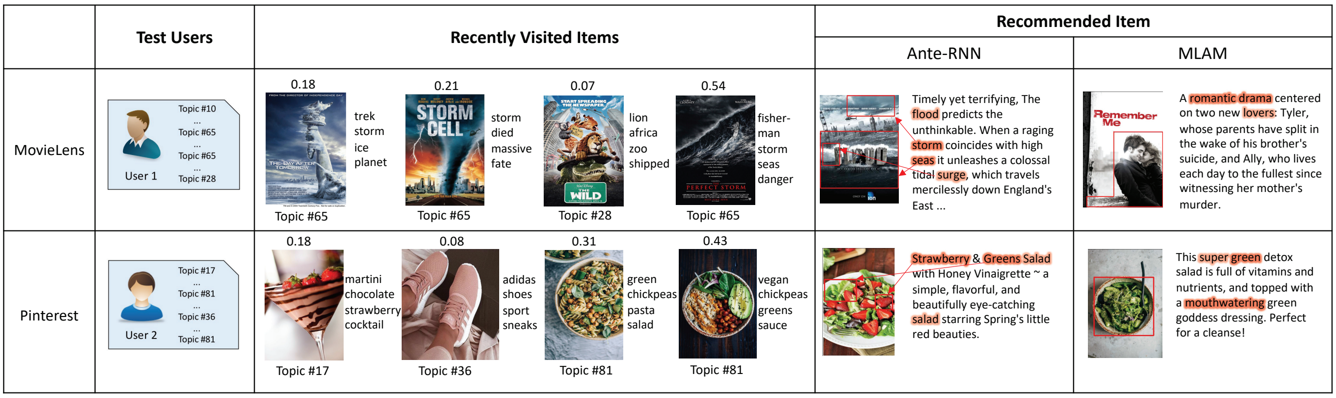
On Pinterest dataset, our model demonstrates the ability of



(a) User1@ MovieLens

(b) User2@ Pinterest

Figure 8: User's interested topic distribution across different weeks on two datasets.



Timely yet terrifying, The flood predicts the unthinkable. When a raging storm coincides with high seas it unleashes a colossal tidal surge, which travels mercilessly down England's East ...

A romantic drama centered on two new lovers: Tyler, whose parents have split in the wake of his brother's suicide, and Ally, who lives each day to the fullest since witnessing her mother's murder.

Strawberry & Greens Salad with Honey Vinaigrette ~ a simple, flavorful, and beautifully eye-catching salad starring Spring's little red beauties.

This super green detox salad is full of vitamins and nutrients, and topped with a mouthwatering green goddess dressing. Perfect for a cleanse!

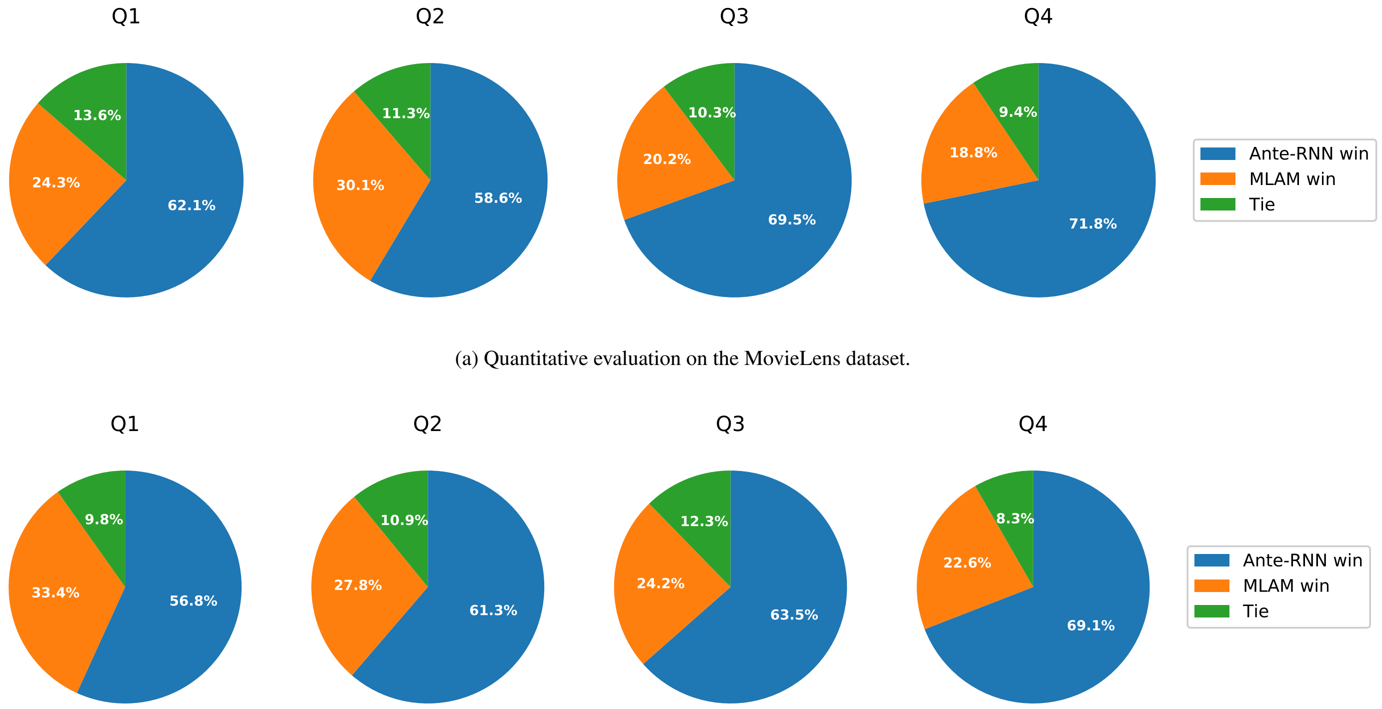
Figure 9: Examples of the visual and textual explanations.

considering both long and short-term interests when recommending items. Specifically, user 2 shows stable interest on topic #81 "Healthy Food" as the yellow line in Figure 8(b) signifies, while she also shows the recent active interest on topic #17 "Cocktail" with cyan dotted line. Consequently, the recommended item shows the combination features on both "Healthy Food" and "Cocktail" with highlighted topic words of *strawberry*, *greens* and *salad*. Meanwhile, *greens* and *strawberry* are marked in image to show the focuses of user's interests. Although MLAM also recommends "Healthy Food" related item, it still prefers the most frequently occurred items and no alignment can be found between visual and textual information.

4.10.2. Quantitative evaluation

To quantitatively evaluate our model's explainability, we conduct crowd-sourcing evaluation by comparing our model with MLAM. Specifically, we select the top-100 most active users from the two datasets separately. For each of the users, we present the image and its corresponding text description of the items that the user previously clicked for the worker to read. The workers are expected to infer this user's personalized preference from these information. Then they will be asked several questions to compare the recommendations and explanations generated by our model and MLAM model. Based on discussions in Tintarev and Masthoff (2007), we carefully designed the survey questions to evaluate different aspects of the recommender algorithm as follows:

- **Q1**: Which recommendation are you more satisfied with?
- **Q2**: Which model could provide you with more ideas about the recommended item?
- **Q3**: Which recommended item are you more likely to



(a) Quantitative evaluation on the MovieLens dataset.

(b) Quantitative evaluation on the Pinterest dataset.

Figure 10: Results of the quantitative evaluation.

click after receiving an explanation?

- **Q4**: Based on the recommended items, which model generated explanation could help you know more easily and clearly why we recommend it to you?

For each question, the workers are required to choose from three options (i.e., A:Ante-RNN, B:MLAM, C:Tie). We intend to use Q1, Q2, Q3 to evaluate satisfaction, effectiveness, and persuasiveness of an explainable recommender algorithm, and use Q4 to judge if our attention mechanism is more effective in this problem.

To perform more accurate evaluations, we recruit 3 workers through Amazon Mechanical Turk for each user's case, and one result is valid only when more than 2 workers share the same opinion. Besides, we require the workers to come from an English-speaking country, older than 18 years, and have online entertainment experience for involving a more diverse population of users. The statistical results are shown in Figure 10. From the results, we can see that our proposed model apparently outperforms MLAM in all aspects of user study. Moreover, the results in Q3 and Q4 manifest that the explanations generated by our model's attention weights could promote the persuasiveness and satisfaction of the recommender algorithm, which verifies the effectiveness of our designed attention mechanism.

4.11. Limitations

We demonstrated that the Ante-RNN model is able to generate both multi-modal and adaptive explanations with recommendation performance comparable to the state-of-the-art methods (Table 4). Yet there are still some limitations: 1) Ante-RNN uses the Faster R-CNN model in conjunction with ResNet-101 pre-trained by Anderson et al. (2018) to learn image region representations. However, not all the images in the dataset are regular and easy to distinguish. Some of them were graffiti, selfies, or even just screenshots of smart phones. Simply adopting a pre-trained weight may cause deviations and inaccurate image-text matching. Moreover, the named entities that are involved in the images cannot be well aligned with text. For example, the ship in the first movie poster of Figure 1 cannot be aligned to "TITANIC" in its corresponding text description. Designing a fine-tuning strategy for the pre-trained model and incorporating knowledge graph into image-text alignment may help with the problem and such is left as a matter for future work. 2) Due to the gating mechanism of recurrent neural networks, our model cannot provide users with a direct and meaningful way to correct the recommendation process if they are unsatisfied with the results. Developing recommendation approaches that are more scrutable would be an interesting research topic and needs to be addressed in future works.

5. Conclusion

User preferences often evolve over time, and it is essential to model their temporal dynamics for recommendation tasks while providing explanations on them. In this paper, we presented an Attentive Recurrent Neural Network (Ante-RNN) for dynamic personalized recommendations. The proposed model allows combining visual image information with text descriptions for better recommendation. Furthermore, a novel hybrid attention mechanism is introduced to strengthen user's short-term preference modelling and capture user's long-term interest dynamics in a better way. The learned attention weights can in turn help to provide reasonable interpretations on recommendation results. We also explore different fusion methods for multi-modality integration. Extensive experiments on two real-world large scale datasets verify that our model can not only provide competitive recommendation performance, but also provide reasonable visual aligned with textual explanations for the recommended items.

One future direction is to enrich user profiles from multi-sources to alleviate cold-start issues for recommendations. Second, we consider other context information, such as time, location, and user sentiments, to further improve our explainability. We also plan to integrate other multi-modalities for a better recommendation performance.

Acknowledgement

This work was supported by the Research Council of Norway (grant number 245469).

References

Ai, Q., Azizi, V., Chen, X., Zhang, Y., 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation. Algorithms 11, 137.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086.

Bao, Y., Fang, H., Zhang, J., 2014. Topicmf: Simultaneously exploiting ratings and reviews for recommendation, in: Twenty-Eighth AAAI conference on artificial intelligence.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y., 2010. Theano: A cpu and gpu math compiler in python, in: Proc. 9th Python in Science Conf, pp. 3–10.

Bucak, S.S., Jin, R., Jain, A.K., 2014. Multiple kernel learning for visual object recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 1354–1369.

Chen, C., Zhang, M., Liu, Y., Ma, S., 2018a. Neural attentional rating regression with review-level explanations, in: Proceedings of the 2018 World Wide Web Conference on World Wide Web, International World Wide Web Conferences Steering Committee. pp. 1583–1592.

Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P., 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM. pp. 163–171.

Chen, X., Qin, Z., Zhang, Y., Xu, T., 2016. Learning to rank features for recommendation over multiple categories, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM. pp. 305–314.

Chen, X., Xu, H., Zhang, Y., Tang, J., Cao, Y., Qin, Z., Zha, H., 2018b. Sequential recommendation with user memory networks, in: Proceedings of the eleventh ACM international conference on web search and data mining, ACM. pp. 108–116.

Chen, X., Zhang, Y., Xu, H., Cao, Y., Qin, Z., Zha, H., 2018c. Visually explainable recommendation. CoRR abs/1801.10288. arXiv:1801.10288.

Cheng, Z., Ding, Y., He, X., Zhu, L., Song, X., Kankanhalli, M.S., 2018. Aˆ 3ncf: An adaptive aspect attention model for rating prediction., in: IJCAI, pp. 3748–3754.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar. pp. 1724–1734.

Cui, Q., Wu, S., Liu, Q., Zhong, W., Wang, L., 2018. Mv-rnn: A multi-view recurrent neural network for sequential recommendation. IEEE Transactions on Knowledge and Data Engineering .

Ding, Y., Li, X., 2005. Time weight collaborative filtering, in: Proceedings of the 14th ACM international conference on Information and knowledge management, ACM. pp. 485–492.

Gao, Y., Chen, J., Zhu, J., 2016. Streaming gibbs sampling for LDA model. CoRR abs/1601.01142. URL: `http://arxiv.org/abs/1601.01142`, `arXiv:1601.01142`.

Geng, X., Zhang, H., Bian, J., Chua, T.S., 2015. Learning image and user features for recommendation in social networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4274–4282.

Grefenstette, E., Hermann, K.M., Suleyman, M., Blunsom, P., 2015. Learning to transduce with unbounded memory, in: Advances in Neural Information Processing Systems, pp. 1828–1836.

Gu, Y., Li, X., Huang, K., Fu, S., Yang, K., Chen, S., Zhou, M., Marsic, I., 2018a. Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder, in: 2018 ACM Multimedia Conference on Multimedia Conference, ACM. pp. 537–545.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I., 2018b. Hybrid attention based multimodal network for spoken language classification, in: Proceedings of the 27th International Conference on Computational Linguistics, pp. 2379–2390.

Guan, Y., Wei, Q., Chen, G., 2019. Deep learning based personalized recommendation with multi-view information integration. Decision Support Systems .

He, R., McAuley, J., 2016. Vbpr: visual bayesian personalized ranking from implicit feedback, in: Thirtieth AAAI Conference on Artificial Intelligence.

Herlocker, J.L., Konstan, J.A., Riedl, J., 2000. Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, ACM. pp. 241–250.

Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D., 2015. Session-based recommendations with recurrent neural networks. CoRR abs/1511.06939.

Hu, L., Jian, S., Cao, L., Chen, Q., 2018. Interpretable recommendation via attraction modeling: Learning multilevel attractiveness over multimodal movie contents., in: IJCAI, pp. 3400–3406.

Huang, J., Zhao, W.X., Dou, H., Wen, J.R., Chang, E.Y., 2018. Improving sequential recommendation with knowledge-enhanced memory networks, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM. pp. 505–514.

Järvelin, K., Kekäläinen, J., 2000. Ir evaluation methods for retrieving highly relevant documents, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM. pp. 41–48.

Kiros, R., Salakhutdinov, R., Zemel, R.S., 2014. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 .

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R., 2016. Ask me anything: Dynamic memory networks for natural language processing, in: International Conference on Machine Learning, pp. 1378–1387.

Lee, K.H., Chen, X., Hua, G., Hu, H., He, X., 2018. Stacked cross attention for image-text matching, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 201–216.

Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., Ma, J., 2017. Neural attentive session-based recommendation, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM. pp. 1419–1428.

Lian, J., Zhang, F., Xie, X., Sun, G., 2018. Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach., in: IJCAI, pp. 3805–3811.

Liang, H., Wang, H., Wang, J., You, S., Sun, Z., Wei, J.M., Yang, Z., 2018. Jtav: Jointly learning social media content representation by fusing textual, acoustic, and visual features. arXiv preprint arXiv:1806.01483 .

Ling, G., Lyu, M.R., King, I., 2014. Ratings meet reviews, a combined approach to recommend, in: Proceedings of the 8th ACM Conference on Recommender systems, ACM. pp. 105–112.

Liu, Q., Wu, S., Wang, L., 2017. Multi-behavioral sequential prediction with recurrent log-bilinear model. IEEE Transactions on Knowledge and Data Engineering 29, 1254–1267.

McAuley, J., Leskovec, J., 2013. Hidden factors and hidden topics: understanding rating dimensions with review text, in: Proceedings of the 7th ACM conference on Recommender systems, ACM. pp. 165–172.

McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A., 2015. Image-based recommendations on styles and substitutes, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM. pp. 43–52.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., Morency, L.P., 2016. Deep multimodal fusion for persuasiveness prediction, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM. pp. 284–288.

Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y., 2016. Jointly modeling embedding and translation to bridge video and language, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4594–4602.

Pei, W., Yang, J., Sun, Z., Zhang, J., Bozzon, A., Tax, D.M., 2017. Interacting attention-gated recurrent networks for recommendation, in: Proceedings of the 2017 ACM on conference on information and knowledge management, ACM. pp. 1459–1468.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P., 2017. Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 873–883.

Poria, S., Chaturvedi, I., Cambria, E., Hussain, A., 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: 2016

IEEE 16th international conference on data mining (ICDM), IEEE. pp. 439–448.

Powers, D.M., 2011. Evaluation: from precision, recall and f-measure to roc informedness, markedness and correlation .

Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L., 2009. Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, AUAI Press. pp. 452–461.

Rendle, S., Freudenthaler, C., Schmidt-Thieme, L., 2010. Factorizing personalized markov chains for next-basket recommendation, in: Proceedings of the 19th international conference on World wide web, ACM. pp. 811–820.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45, 2673–2681.

Seo, S., Huang, J., Yang, H., Liu, Y., 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM. pp. 297–305.

Song, Y., Elkahky, A.M., He, X., 2016. Multi-rate deep learning for temporal recommendation, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM. pp. 909–912.

Sukhbaatar, S., Weston, J., Fergus, R., et al., 2015. End-to-end memory networks, in: Advances in neural information processing systems, pp. 2440–2448.

Sun, M., Li, F., Zhang, J., 2018. A multi-modality deep network for cold-start recommendation. Big Data and Cognitive Computing 2, 7.

Tan, Y., Zhang, M., Liu, Y., Ma, S., 2016. Rating-boosted latent topics: Understanding users and items with ratings and reviews., in: IJCAI, pp. 2640–2646.

Tay, Y., Luu, A.T., Hui, S.C., 2018. Multi-pointer co-attention networks for recommendation, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM. pp. 2309–2318.

Tintarev, N., Masthoff, J., 2007. A survey of explanations in recommender systems, in: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop, IEEE Computer Society, Washington, DC, USA. pp. 801–810.

Tintarev, N., Masthoff, J., 2011. Designing and evaluating explanations for recommender systems, in: Recommender systems handbook. Springer, pp. 479–510.

Wang, C., Blei, D.M., 2011. Collaborative topic modeling for recommending scientific articles, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 448–456.

Wang, G., Yan, J., Qin, Z., 2018a. Collaborative and attentive learning for personalized image aesthetic assessment., in: IJCAI, pp. 957–963.

Wang, H., Zhang, F., Xie, X., Guo, M., 2018b. Dkn: Deep knowledge-aware network for news recommendation, in: Proceedings of the 2018 World Wide Web Conference on World Wide Web, International World Wide Web Conferences Steering Committee. pp. 1835–1844.

Wang, N., Wang, H., Jia, Y., Yin, Y., 2018c. Explainable recommendation via multi-task learning in opinionated text data, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM. pp. 165–174.

Wang, P., Guo, J., Lan, Y., Xu, J., Wan, S., Cheng, X., 2015. Learning hierarchical representation model for nextbasket recommendation, in: Proceedings of the 38th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM. pp. 403–412.

Wang, S., Wang, Y., Tang, J., Shu, K., Ranganath, S., Liu, H., 2017. What your images reveal: Exploiting visual contents for point-of-interest recommendation, in: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee. pp. 391–400.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.P., 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems 28, 46–53.

Wörtwein, T., Scherer, S., 2017. What really mattersan information gain analysis of questions and reactions in automated ptsd screenings, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE. pp. 15–20.

Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T., 2016. A dynamic recurrent model for next basket recommendation, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM. pp. 729–732.

Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P., 2017. Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark. pp. 1103–1114. URL: `https://www.aclweb.org/anthology/D17-1115`, doi:`10.18653/v1/D17-1115`.

Zhang, S., Zhang, S., Huang, T., Gao, W., Tian, Q., 2018. Learning affective features with a hybrid deep model for audio–visual emotion recognition. IEEE Transactions on Circuits and Systems for Video Technology 28, 3030–3043.

Zhang, W., Wang, J., 2016. Integrating topic and latent factors for scalable personalized review-based rating prediction. IEEE Transactions on Knowledge and Data Engineering 28, 3013–3027.

Zhang, Y., Ai, Q., Chen, X., Croft, W.B., 2017. Joint representation learning for top-n recommendation with heterogeneous information sources, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM. pp. 1449–1458.

Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma, S., 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM. pp. 83–92.

Zhu, Y., Li, H., Liao, Y., Wang, B., Guan, Z., Liu, H., Cai, D., 2017. What to do next: Modeling user behaviors by time-lstm., in: IJCAI, pp. 3602–3608.