

CATA++: A Collaborative Dual Attentive Autoencoder Method for Recommending Scientific Articles^{*}

Meshal Alfarhood^{a,*}, Jianlin Cheng^{a,b}

^aDepartment of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

^bInformatics Institute, University of Missouri, Columbia, MO 65211, USA

ARTICLE INFO

Keywords:

Recommender systems
Collaborative filtering
Matrix factorization
Sparsity problem
Attention mechanism
Autoencoder
Deep learning

ABSTRACT

Recommender systems today have become an essential component of any commercial website. Collaborative filtering approaches, and Matrix Factorization (MF) techniques in particular, are widely used in recommender systems. However, the natural data sparsity problem limits their performance where users generally interact with very few items in the system. Consequently, multiple hybrid models were proposed recently to optimize MF performance by incorporating additional contextual information in its learning process. Although these models improve the recommendation quality, there are two primary aspects for further improvements: (1) multiple models focus only on some portion of the available contextual information and neglect other portions; (2) learning the feature space of the side contextual information needs to be further enhanced.

In this paper, we propose a Collaborative Dual Attentive Autoencoder (CATA++) for recommending scientific articles. CATA++ utilizes an article's content and learns its latent space via two parallel autoencoders. We use attention mechanism to capture the most pertinent part of information in making more relevant recommendations. Comprehensive experiments on three real-world datasets have shown that our dual-way learning strategy has significantly improved the MF performance in comparison with other state-of-the-art MF-based models according to various experimental evaluations. The source code of our methods is available at: <https://github.com/jianlin-cheng/CATA>.

1. Introduction

The amount of data created in the last few years is overwhelming. Interestingly, the data volume grows exponentially yearly compared to the years before, making the era of big data. This motivates and attracts researchers to utilize this massive data to develop more practical and accurate solutions in most computer science domains. For example, recommender systems (RSs) are primarily a good solution to process big data in order to extract useful information, e.g. users' preferences, to help users with personalized decision making.

Scientific article recommendation is a very common application for RSs. It keeps researchers updated on recent related work in their field. One traditional way to find relevant articles is to go through the references section in other articles. Yet, this traditional approach is biased to heavily cited articles, such that new relevant articles with higher impact have less chance to be found. Another method is to search articles with keywords. Although this technique is popular among researchers, they need to filter out a tremendous number of articles in the searching results to retrieve the most suited articles for them. Besides that, all users get the same searching results with the same keywords, and they are not


personalized results based on the users' personal interests. Thus, recommendation systems are the key solution for such issues to help scientists and researchers find impressive articles and be aware of recent related work.

Collaborative filtering models (CF) are widely successful models applied to RSs. CF models depend typically on users' ratings about items, such that users with similar past ratings are more likely to agree on similar items in the future. Matrix Factorization (MF) is one of the most popular CF techniques for many years and has been widely used in the recommendation literature. Many proposed models are enhanced versions of MF [11, 16, 20, 10, 7]. However, CF models generally rely only on users' past ratings in their learning process, and do not consider other auxiliary information, which has been validated later to improve the quality of recommendations. For that reason, the performance of CF models decreases significantly when users have limited, insufficient amount of ratings data. This problem is also known as the data sparsity problem.

More recently, much efforts have been conducted to include item's information along with the user's ratings data via topic modeling [26, 27, 15]. Collaborative Topic Regression (CTR) [26] for example is composed of Probabilistic Matrix Factorization (PMF) and Latent Dirichlet Allocation (LDA) to utilize both user's ratings and item's reviews to learn their latent features. By doing that, the natural sparsity problem could be alleviated, and these kinds of approaches are called hybrid models. Hybrid models generally are divided into two sub-categories according to how models are trained: loosely coupled models and tightly coupled models [28]. Loosely coupled models train CF and content-based filtering (CBF) models separately, like ensembles, and then

^{*} A preliminary version of this article has been presented at the IEEE ICMLA conference 2019 [1]. However, this submission has substantially extended our previous work by improving the model architecture, and adding extensive experimental contributions in comparison with the conference paper.

*Corresponding author

 may82@missouri.edu (M. Alfarhood); chengji@missouri.edu (J. Cheng)

Cheng)

ORCID(s):

find out the final score based on the scores of the two separated models. On the other hand, the tightly coupled models train both CF and CBF models jointly. In joint training, both models are cooperating with each other to calculate the final score under the same loss function.

Simultaneously, machine learning, and deep learning (DL) in particular, have gained increasing attention in recent years due to how they enhance the way we process big data, and to their capability of modeling complicated data such as texts and images. DL meets recommendation systems the last few years and has shown superiority over traditional collaborative filtering models. Restricted Boltzmann Machines (RBM) [21] is one of the first works that applies DL for CF recommendations. However, RBM is not deep enough to learn users' tastes from the users' feedback data, and also it does not take side information into consideration. Later on, Collaborative Deep Learning (CDL) [28] became the state-of-the-art method in DL-based RSs. CDL can be viewed as an updated version of CTR [26] by substituting the LDA topic modeling with a Stacked Denoising Autoencoder (SDAE) to learn from item contents. In addition, Deep Collaborative Filtering (DCF) [12] is a similar work that uses a marginalized Denoising Autoencoder (mDA) with PMF. Lately, Collaborative Variational Autoencoder (CVAE) [13] has been proposed, which uses a variational autoencoder to handle the item contents, and has shown to have better predictions over CDL.

However, existing recommendation models, such as CDL and CVAE, have two limitations. First, they assume that all parts of their model's contribution are the same in their final predictions. Second, they focus only on some parts of item's information and neglect other parts, which can be also utilized in the recommendation process.

In this work, we propose a deep learning-based model named Collaborative Dual Attentive Autoencoder (CATA++) that has been evaluated on scientific article recommendation's task. We integrate attention technique into our deep feature learning procedure to learn from article's textual information, such as title, abstract, tags, and citations, to enhance the recommendation quality. The compressed low-dimensional representation learned by each unsupervised model is incorporated then into matrix factorization approach for our ultimate recommendation. To demonstrate the capability of our proposed model to generate relevant recommendations, we conduct inclusive experiments on three real-world datasets, taken from Citeulike¹ website, to evaluate CATA++ in comparison with multiple recent MF-based models. The experimental results have proved that our model can extract more constructive information from the article's contextual data that leads into better recommendation performance where the data sparsity is extremely high.

The main contributions of this work are summarized in the following points:

- We introduce CATA++, a Collaborative Dual Attentive Autoencoder that has been evaluated on recom-

mending scientific articles. We employ attention mechanism into our model to work between the encoder and the decoder such that only relevant parts of the information can contribute more in representing the item content. This item's representation helps in finding the similarities between articles.

- We exploit more article content into our deep feature learning process. To the best of our knowledge, our model is the first model that utilizes all article content including title, abstract, tags, and citations between articles all together in one model by coupling two attentive autoencoder networks. The latent features learned by each network are then integrated into a matrix factorization method for our ultimate recommendations.
- We evaluate our model on three real-world datasets. We compare the performance of our proposed model with five baselines. CATA++ achieves superior performance when the data sparsity is extremely high.

The remainder of this paper is organized as follows. We explain some preliminaries in Section 2. Our model, CATA++, has been demonstrated in depth in Section 3. The experimental results of our model against the state-of-the-art models have been discussed thoroughly in Section 4. We then conclude our work in Section 5.

2. Preliminaries

Our work is designed and evaluated on recommendations with implicit feedback. In this section, we describe the well-known collaborative filtering approach, Matrix Factorization, for implicit feedback problems, and then followed by the definition of the attention mechanism and its related work.

2.1. Matrix Factorization

Matrix Factorization (MF) [11] is the most popular CF method, mainly due to its simplicity and efficiency. The idea behind MF is to decompose the user-item matrix, $R \in \mathbb{R}^{n \times m}$, into two lower-dimensional matrices, $U \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{m \times d}$, such that the inner product of U and V will approximate the original matrix R , where d is the dimension of the latent factors.

$$R \approx U \cdot V^T \quad (1)$$

MF optimizes values of U and V by minimizing the sum of the squared difference between actual values and model predictions with adding two regularization terms as follows:

$$\mathcal{L} = \sum_{i,j \in R} \frac{I_{ij}}{2} (r_{ij} - u_i v_j^T)^2 + \frac{\lambda_u}{2} \|u_i\|^2 + \frac{\lambda_v}{2} \|v_j\|^2 \quad (2)$$

where I_{ij} is an indicator function that equals 1 if user i has rated item j , and 0 otherwise. Also, $\|U\|$ and $\|V\|$ are the Euclidean norms, and λ_u, λ_v are two regularization terms

¹www.citeulike.org

preventing the values of U and V from being too large to avoid the model overfitting.

Explicit data such as ratings, r_{ij} , are not regularly available. Therefore, Weighted Regularized Matrix Factorization (WRMF) [7] introduced two modifications to the previous objective function to make it work for implicit feedback. The optimization process in this case runs through all user-item pairs with different confidence levels assigned to each pair as the following:

$$\mathcal{L} = \sum_{i,j \in R} \frac{c_{ij}}{2} (p_{ij} - u_i v_j^T)^2 + \frac{\lambda_u}{2} \|u_i\|^2 + \frac{\lambda_v}{2} \|v_j\|^2 \quad (3)$$

where p_{ij} is the user preference score that has a value of 1 when user_{*i*} and item_{*j*} have interaction, and 0 otherwise. c_{ij} is a confidence variable where its value shows how confident the user like the item. In general, $c_{ij} = a$ when $p_{ij} = 1$, and $c_{ij} = b$ when $p_{ij} = 0$, such that $a > b > 0$.

2.2. Attention mechanism

The idea of the attention mechanism is motivated by the human vision system and how our eyes pay attention and focus to a specific part of an image, or specific words in a sentence, for example. In the same way, attention in deep learning can be described simply as a vector of weights to show the importance of the input elements. Thus, the intuition behind attention is that not all parts of the input are equally significant, i.e., only few parts are significant for the model. Attention was initially designed for image classification task [17], and then successfully applied in natural language processing (NLP) for machine translation task [3] when the input and the output may have different lengths.

Attention has also been successfully applied in different recommendation tasks [9, 14, 25, 29, 23, 4]. For example, MPCN [25] is a multi-pointer co-attention network that takes user and item reviews as input, and then extracts the most informative reviews that contribute more in predictions. Also, D-Attn [23] uses a convolutional neural network with dual attention (local and global attention) to represent the user and the item latent representations similarly like matrix factorization approach. Moreover, NAIS [4] employs attention network to distinguish items in a user profile, which have more influential effects in the model predictions.

3. Methodology

In this section, we illustrate our proposed model in depth. The intuition behind our model is to learn the latent factors of items in PMF with the use of available side textual contents using two parallel attentive unsupervised learning models that can catch more plentiful information from the available data. The architecture of our model is displayed in Figure 1.

Before going through our model, we first define our recommendation problem. The recommendation problem with implicit data is usually formulated as follows:

$$p_{ui} = \begin{cases} 1, & \text{if there is user-item interaction} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Table 1

A summary of notations description.

Notation	Meaning
n	Number of users
m	Number of articles
d	Dimension of the latent factors
R	User-article matrix
U	Latent factors of users
V	Latent factors of articles
C	Confidence matrix
P	Users preferences matrix
X_j	Article's side information, i.e., title and abstract
T_j	Article's side information, i.e., tags and citations
$\theta(X_j)$	Mapping function for input X_j of the autoencoder
$\gamma(T_j)$	Mapping function for input T_j of the autoencoder
Z_j	Compressed representation of X_j
Y_j	Compressed representation of T_j
λ_u	Regularization parameter of users
λ_v	Regularization parameter of articles

where the ones show positive feedback, and the zeros show missing values. Therefore, a zero value could mean either negative feedback, or the user is unaware of that item. Generally, implicit feedback can be obtained from the users' behaviors such as users' clicks and bookmarks, because explicit feedback such as users' ratings occasionally are not available due to the difficulty of obtaining users' explicit opinions. Implicit feedback is widely used by ranking prediction models [7, 18, 19]. Ranking prediction works by suggesting a list of items to a user and ranking them based on the user preferences. Even though our model has been applied to a ranking prediction problem with implicit feedback data, it could be used for a rating prediction problem with explicit feedback data as well by altering the final loss function.

We now clarify each part of our model individually in the following sections. Table 1 summarizes all the notations used in this paper to describe our approach.

3.1. The attentive autoencoder

Autoencoder [5] is an unsupervised learning neural network that is useful for compressing high-dimensional input data into a lower representation while preserving the abstract information of the data. The network is composed of two main parts, which are the encoder and the decoder. The encoder takes the input and encodes it through multiple hidden layers into a lower-dimensional compressed representation, Z_j . The encoding function can be formulated as $Z_j = f(X_j)$. On the other hand, the decoder can be used then to reconstruct the estimated input, \hat{X}_j , from the latent space representation. The decoder function can be formulated as $\hat{X}_j = f(Z_j)$. Each of the encoder and the decoder usually consist of the same number of hidden layers. The output of each hidden layer is computed as follows:

$$h^{(\ell)} = \sigma(h^{(\ell-1)}W^{(\ell)} + b^{(\ell)}) \quad (5)$$

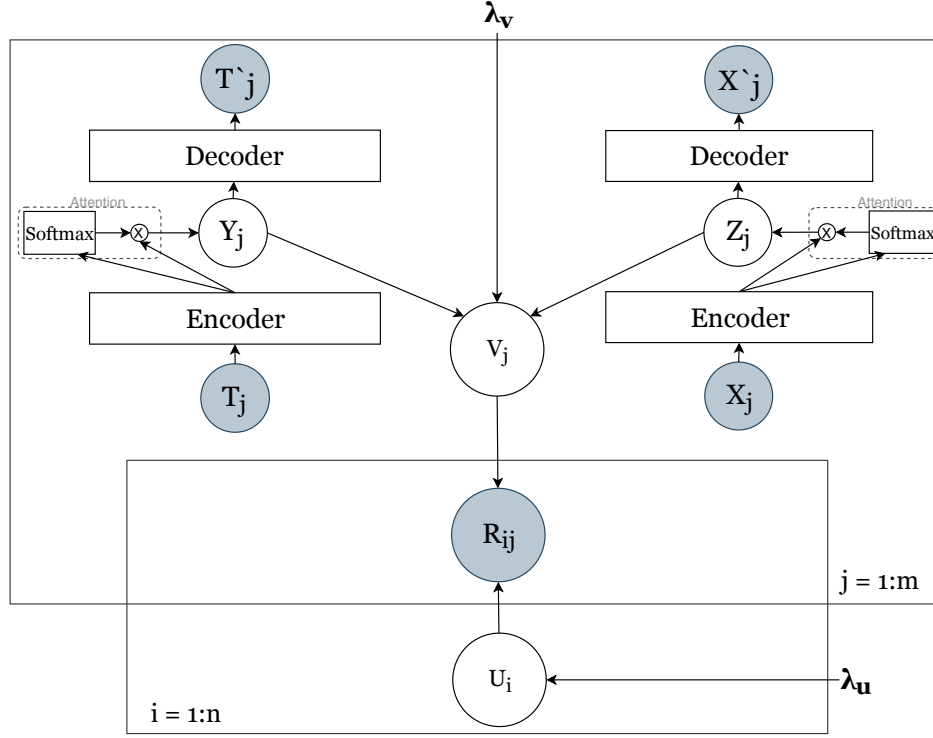


Figure 1: Collaborative Dual Attentive Autoencoder (CATA++) architecture.

where (ℓ) is the layer number, W are the weights matrix, b is the bias vector, and σ is a non-linear activation function. We use the Rectified Linear Unit (ReLU) as the activation function.

Our model takes two inputs from article information: $X_j = \{x^1, x^2, \dots, x^s\}$ and $T_j = \{t^1, t^2, \dots, t^g\}$ where x^i and t^i are values between $[0, 1]$, s is the vocabulary size of the articles' titles and abstracts, and g is the vocabulary size of the articles' tags. In other words, the inputs of the network in our experiments are two normalized bag-of-words histograms of filtered vocabularies of the articles' textual data.

Batch normalization (BN) [8] has been proven to be a proper solution for the internal covariant shift problem, where the layer's input distribution in deep neural networks changes over time of training, and causes difficulty to train the model. In addition, BN can work as a regularization procedure like Dropout [24] in deep neural networks. Accordingly, we apply a batch normalization layer after each hidden layer in our autoencoder to obtain a stable distribution of the output of each layer, which has a useful effect eventually on the model accuracy.

Furthermore, we use the idea of attention mechanism to work between the encoder and the decoder such that only relevant parts of the encoder output are selected for the input reconstruction. We first calculate the scores as the probability distribution of the encoder's output using the *softmax(.)* function.

$$f(z_c) = \frac{e^{z_c}}{\sum_d e^{z_d}} \quad (6)$$

The probability distribution and the encoder output are then multiplied using element-wise multiplication function to get Z_j .

We use the dual attentive autoencoder to pretrain all items' contextual information and then integrate the two compressed representations, Z_j and Y_j , in computing the latent factors of items, V_j , from the matrix factorization method. The dimension space of Z_j , Y_j , and V_j are set to be equal to each other. Finally, we adopt the binary cross-entropy (Equation 7) as the loss function we want to minimize for each attentive autoencoder model as:

$$\mathcal{L} = - \sum_k (y_k \log(p_k) - (1 - y_k) \log(1 - p_k)) \quad (7)$$

where y_k corresponds to the correct labels, and p_k corresponds to the predicted values.

The value of p that minimizes the previous loss function the most is when $p = y$, which makes it fit for our autoencoder. To verify that, taking the derivative of the loss function to respect to p results into:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p} &= -y \left(\frac{1}{p} \right) - (1 - y) \left(\frac{-1}{1 - p} \right) \\ \frac{-y}{p} + \frac{1 - y}{1 - p} &= 0 \\ -y(1 - p) + (1 - y)p &= 0 \\ -y + yp + p - yp &= 0 \\ -y + p &= 0 \\ p &= y \end{aligned} \quad (8)$$

3.2. Probabilistic matrix factorization

Probabilistic Matrix Factorization (PMF) [16] is a probabilistic linear model where the prior distributions of the latent factors and users' preferences are drawn from Gaussian normal distribution. In our previous model, CATA [1], we integrate items contents trained through a single attentive autoencoder into PMF. The objective function for CATA was defined as:

$$\mathcal{L} = \sum_{i,j \in R} \frac{c_{ij}}{2} (p_{ij} - u_i v_j^T)^2 + \frac{\lambda_u}{2} \|u_i\|^2 + \frac{\lambda_v}{2} \|v_j - \theta(X_j)\|^2 \quad (9)$$

where $\theta(X_j) = \text{Encoder}(X_j) = Z_j$ such that $\theta(X_j)$ works as the Gaussian prior information to v_j .

However, CATA++ exploits more items contents trained through two parallel attentive autoencoder into PMF. Therefore, the objective function has been changed slightly to become:

$$\mathcal{L} = \sum_{i,j \in R} \frac{c_{ij}}{2} (p_{ij} - u_i v_j^T)^2 + \frac{\lambda_u}{2} \|u_i\|^2 + \frac{\lambda_v}{2} \|v_j - (\theta(X_j) + \gamma(T_j))\|^2 \quad (10)$$

where $\gamma(T_j) = \text{Encoder}(T_j) = Y_j$.

To determine what values of user and item vectors that minimize the previous objective function (Equation 10), we first take the derivative of \mathcal{L} with respect to u_i .

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial u_i} &= - \sum_j c_{ij} (p_{ij} - u_i v_j^T) v_j + \lambda_u u_i \\ 0 &= -C_i (P_i - u_i V^T) V + \lambda_u u_i \\ 0 &= -C_i V P_i + C_i V u_i V^T + \lambda_u u_i \\ V C_i P_i &= u_i V C_i V^T + \lambda_u u_i \\ V C_i P_i &= u_i (V C_i V^T + \lambda_u I) \\ u_i &= V C_i P_i (V C_i V^T + \lambda_u I)^{-1} \\ u_i &= (V C_i V^T + \lambda_u I)^{-1} V C_i P_i \end{aligned} \quad (11)$$

where I is the identity matrix.

Similarly, taking the derivative of \mathcal{L} with respect to v_j leads to:

$$v_j = (U C_j U^T + \lambda_v I)^{-1} U C_j P_j + \lambda_v (\theta(X_j) + \gamma(T_j)) \quad (12)$$

Vectors u_i and v_j are updated using Alternating Least Squares (ALS) optimization method where it iteratively optimizes U while V is fixed and vice versa. This optimization process is repeated until the model converges.

3.3. Prediction

After our model has been trained, and the latent factors of users and articles, U and V , are identified, we calculate our model's prediction scores of user i and each article as the dot product of vector u_i with all vectors in V as $scores_i = u_i V^T$. Then, we sort all articles based on our model predication scores in descending order, and then recommend the top- K articles for that user i . We go through all users in U in our evaluation and report the average performance among all users. The overall process of our approach is illustrated in Algorithm 1.

Algorithm 1: CATA++ algorithm

```

1 pre-train first autoencoder with input  $X$ ;
2 pre-train second autoencoder with input  $T$ ;
3  $Z \leftarrow \theta(X)$ ;
4  $Y \leftarrow \gamma(T)$ ;
5  $U, V \leftarrow$  Initialize with random values;
6 while <NOT converge> do
7   for <each user  $i$ > do
8      $u_i \leftarrow$  update using Equation 11;
9   end for
10  for <each article  $j$ > do
11     $v_j \leftarrow$  update using Equation 12;
12  end for
13 end while
14 for <each user  $i$ > do
15    $scores_i \leftarrow u_i V^T$ ;
16   sort( $scores_i$ ) in descending order;
17 end for
18 Evaluate the top- $K$  recommendations;

```

4. Experiments

In this section, we conduct extensive experiments aiming to answer the following research questions:

- **RQ1:** How does our proposed model, CATA++, perform against the state-of-the-art methods? Show quantitative and qualitative analysis.
- **RQ2:** Are both autoencoders (left and right) cooperating with each other for the ultimate recommendation performance?
- **RQ3:** What is the impact of different hyper-parameters tuning (e.g. dimension of features' latent space, number of layers inside each encoder and decoder, and regularization terms λ_u and λ_v) on the performance of our model?

We first describe the datasets, evaluation methodology, baselines, followed by the experimental results answering the previous research questions.

4.1. Datasets

We use three real-world, scientific article datasets to evaluate our model against the state-of-the-art models. All three datasets are collected from Citeulike website. Citeulike was a web service that let users to create their own library of academic publications.

First, Citeulike-a dataset is collected by [26] and it has 5,551 users, 16,980 articles, 204,986 user-article interaction pairs, 46,391 tags, and 44,709 citations between articles. Citations between articles are taken from Google Scholar². The sparseness of this dataset is extremely high with only around 0.22% of the user-article matrix having interactions. Each user has at least 10 articles in his library. On average, each

²<https://scholar.google.com>

Table 2
Description of Citeulike datasets.

Dataset	#users	#articles	#pairs	#tags	#citations	sparsity%
Citeulike-a	5,551	16,980	204,986	46,391	44,709	99.78%
Citeulike-t	7,947	25,975	134,860	52,946	32,565	99.93%
Citeulike-2004-2007	3,039	210,137	284,960	75,721	–	99.95%

user has 37 articles in his library and each article has been added to 12 users' libraries.

Second, Citeulike-t dataset is collected by [27] and it has 7,947 users, 25,975 articles, 134,860 user-article interaction pairs, 52,946 tags, and 32,565 citations between articles. This dataset is actually more sparse than the first one with only 0.07% of the user-article matrix having interactions. Each user has at least 3 articles in his library. On average, each user has 17 articles in his library in this dataset and each article has been added to 5 users' libraries.

Third, Citeulike-2004-2007 dataset is three times bigger than the previous ones with regards to the user-article matrix. The data values in this dataset are extracted between the period of 11-04-2004 until 12-31-2007. It is collected by [2] and it has 3,039 users, 210,137 articles, 284,960 user-article interaction pairs, and 75,721 tags. The tags are single-word keywords that have been generated by Citeulike users when they add an article to their library. Also, it is worth noticing that citations data is not available in this dataset. This dataset is even the most sparse dataset in this experiment with sparsity equal to 99.95%. Each user has at least 10 articles in his library. On average, each user has 94 articles in his library and each article has been added only to 1 user library. Also, this dataset poses a scalability challenge for recommender systems because of its size. Summarized statistics of the datasets are shown in Table 2.

Figure 2 shows the ratio of articles that have been added to 5 users' libraries or less. For example, 15%, 77%, and 99% of the articles in Citeulike-a, Citeulike-t, and Citeulike-2004-2007 respectively are added to 5 users' libraries or less. Moreover, only 1% of the articles in Citeulike-a have been added only to 1 user library, while the rest of the articles have been added to more than this number. On the contrary, 13%, and 77% of the articles in Citeulike-t and Citeulike-2004-2007 have been added only to 1 user library. This proves the sparseness of the data with regards to articles as we go from one dataset to another.

Title and abstract of each article are given. The average number of words per article in both title and abstract after our text preprocessing is 67 words in Citeulike-a, 19 words in Citeulike-t, and 55 words in Citeulike-2004-2007. We follow the same techniques as the state-of-the-art models [28, 26, 13] to preprocess our textual content. First, the title and the abstract of each article are combined together and then preprocessed such that stop words are removed. After that, top-N distinct words based on the TF-IDF measurement [22] are picked out. 8,000 distinct words are selected for Citeulike-a, 20,000 distinct words are selected for Citeulike-

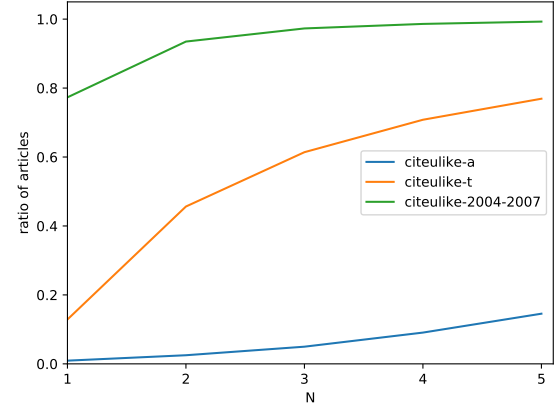


Figure 2: Ratio of articles that have been added to $\leq N$ users' libraries.

t, and 19,871 distinct words are selected for Citeulike-2004-2007 to form the bag-of-words histograms, which are then normalized into values between 0 and 1 based on the vocabularies' occurrences.

Similarly, we preprocess the tags information such that tags assigned to less than 5 articles are removed, and thus we get 7,386 and 8,311 tags in total for Citeulike-a and Citeulike-t, respectively. For Citeulike-2004-2007 dataset, we only kept tags that are assigned to more than 10 articles, and that results in 11,754 tags in total for this dataset. After that, we create a matrix of bag-of-words histogram, $Q \in \mathbb{R}^{m \times g}$, to represent the article-tag relationship, with m being the number of articles, and g being the number of tags. This matrix is filled with ones and zeros such that:

$$q_{at} = \begin{cases} 1, & \text{if tag}_t \text{ is assigned to article}_a \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

Also, citations between articles are integrated in this matrix such that if $article_x$ cites $article_y$, then all the ones in vector q_y of the original matrix are copied into vector q_x . We do that to capture the article-article relationship.

4.2. Evaluation methodology

We follow the state-of-the-art techniques [13, 28, 27] to generate the training and testing sets. For each dataset, we create two versions of the dataset for sparse and dense settings. In total, 4 dataset cases are used in our evaluations. To form the sparse ($P = 1$) and the dense ($P = 10$) datasets, P articles are randomly selected from each user library to generate the training set while the remaining articles from each

user library are used to generate the testing set. As a result, when $P = 1$, only 2.7%, 5.9%, and 1.1% of the data entries are used to generate the training set in Citeulike-a, Citeulike-t, and Citeulike-2004-2007 respectively. Similarly, 27.1%, 39.6%, and 10.7% of the data entries are used to generate the training set when $P = 10$.

In our evaluations, we repeat the data splitting 4 times with randomly different splits of training and testing set. We use one split as a validation experiment to find the optimal parameters for each of model, while the other three splits are used to report the average performance of our model against the baselines.

We use recall and normalized Discounted Cumulative Gain (nDCG) as our evaluation metrics to test how our model performs. Recall is usually used to evaluate recommender systems with implicit feedback. However, precision is not favorable to use with implicit feedback because the zero value in the user-article matrix means either the user is not interested in the article (negative feedback), or the user is not aware of the existing of this article. However, precision metric only treats each zero value as a negative feedback.

Recall per user can be measured using the following formula:

$$\text{recall}@K = \frac{\text{Relevant Articles} \cap K \text{ Recommended Articles}}{\text{Relevant Articles}} \quad (14)$$

where K is set manually in the experiment and it represents the top- K articles of each user. We set $K = 10, 50, 100, 150, 200, 250$, and 300 in our evaluations. The overall recall can be calculated as the average recall among all users. If K equals to the number of articles in the dataset, recall basically equals 1.

Recall, however, does not take into account the ranking of articles within the top- K recommendations as long as they are in the top- K list. However, nDCG does. nDCG shows the capability of the recommendation engine to recommend articles at the top of the ranking list. Articles in higher ranked positions have more value than others. nDCG among all users can be measured using the following equation:

$$\text{nDCG}@K = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{\text{DCG}@K}{\text{IDCG}@K} \quad (15)$$

such that:

$$\begin{aligned} \text{DCG}@K &= \sum_{i=1}^K \frac{\text{rel}(i)}{\log_2(i+1)} \\ \text{IDCG}@K &= \sum_{i=1}^{\min(R,K)} \frac{1}{\log_2(i+1)} \end{aligned} \quad (16)$$

where $|U|$ is the total number of users, i is the rank of the top- K articles recommended by the model, R is the number of the relevant articles, and $\text{rel}(i)$ is an indicator function that outputs 1 if the article at rank i is a relevant article, and 0 otherwise.

4.3. Baselines

We evaluate our approach against the following baselines:

- **POP**: Popular predictor is a non-personalized recommender system. It recommends the most popular articles in the training set to all users. It is widely used as the baseline for personalized recommender systems models.
- **CDL**: Collaborative Deep Learning (CDL) [28] is a probabilistic model that jointly models both the user-item matrix and the items content using a stacked denoising autoencoder (SDAE) with a probabilistic matrix factorization (PMF).
- **CVAE**: Collaborative Variational Autoencoder (CVAE) [13] is a similar approach to CDL [28]. However, it uses a variational autoencoder (VAE) instead of SDAE to incorporate the item content into PMF.
- **CVAE++**: We modify the implementation of CVAE [13] to include two variational autoencoders to engage more side information into the model training, like what CATA++ does. As a result of adding another VAE into the model, we change the loss function accordingly such that the loss of the item latent variable becomes: $\mathcal{L}(v) = \lambda_v \sum_j \|v_j - (z_j + y_j)\|_2^2$, where z_j is the latent content variable of the first VAE, and y_j is the latent content variable of the second VAE.
- **CATA**: Collaborative Attentive Autoencoder (CATA) [1] is our preliminary work that uses a single attentive autoencoder (AAE) to train article content, i.e., title and abstract.

Table 3 gives more clarifications about which part of article's data is involved in each model training. As the table shows, only CATA++ and CVAE++ use all the available information for training their model.

Table 4 also reports the best values of λ_u and λ_v for CDL, CVAE, CVAE++, CATA, and CATA++ based on the validation experiment. We use a grid search of the following values $\{0.01, 0.1, 1, 10, 100\}$ to obtain the optimal values. Moreover, for CDL, we set $a=1$, $b=0.01$, $D=50$, $\lambda_n=1000$, and $\lambda_w=0.0001$. Also, we use a 2-layer SDAE network architecture that has a structure of "#Vocabularies-200-50-200-#Vocabularies" to run their code on our datasets. Similarly, for CVAE and CVAE++, we also set $a=1$, $b=0.01$, and $D=50$. A three-layer VAE network architecture that is similar to the structure reported in their paper is used with a structure equivalent to "#Vocabularies-200-100-50-100-200-#Vocabularies". Finally, for CATA and CATA++, we also set $a=1$, $b=0.01$, and $D=50$. A four-layer AAE network architecture in the form of "#Vocabularies-400-200-100-50-100-200-400-#Vocabularies" is used train our models.

4.4. Experimental results

We now answer the research questions that have been previously defined in the beginning of this section.

Table 3

Comparison between all models about which data they use in their model training.

Approach	User-article matrix	Side information			
		Title	Abstract	Tags	Citations
POP	✓	–	–	–	–
CDL	✓	✓	✓	–	–
CVAE	✓	✓	✓	–	–
CVAE++	✓	✓	✓	✓	✓
CATA	✓	✓	✓	–	–
CATA++	✓	✓	✓	✓	✓

Table 4Parameter settings for λ_u and λ_v for CDL, CVAE, CVAE++, CATA, and CATA++ based on the validation experiment.

Approach	Citeulike-a				Citeulike-t				Citeulike-2004-2007			
	Sparse		Dense		Sparse		Dense		Sparse		Dense	
	λ_u	λ_v	λ_u	λ_v	λ_u	λ_v	λ_u	λ_v	λ_u	λ_v	λ_u	λ_v
CDL	0.01	10	0.01	10	0.01	10	0.01	10	0.01	10	0.01	10
CVAE	0.1	10	1	10	0.1	10	0.1	10	0.1	10	0.1	10
CVAE++	0.1	10	0.1	10	0.1	10	0.1	10	1	10	1	10
CATA	10	0.1	10	0.1	10	0.1	10	0.1	10	0.1	10	0.1
CATA++	10	0.1	10	0.1	10	0.1	10	0.1	10	0.1	10	0.1

4.4.1. RQ1

To measure the performance of our model against the baselines, we conduct quantitative and qualitative comparisons to answer this question. Figures 3 and 4 show the performance of the top- K recommendation under the sparse setting, $P = 1$, based on recall and nDCG using all the three datasets. In the same way, Figures 5 and 6 show the performance under the dense setting, $P = 10$, based on recall and nDCG as well.

First, the sparse cases are very critical and challenging for any proposed model since there is less feedback data for training. In the sparse setting where there is only one article in each user's library for the training phase, our model, CATA++, outperforms the state-of-the-art MF-based models in all three datasets in terms of recall and nDCG. More importantly, CATA++ beats the best model among all the baselines, CVAE++, by a wide margin in Citeulike-2004-2007, where it's actually more sparse and contains huge number of articles. This validates the robustness of our model against the data sparsity. Second, in the dense setting where there are more articles in each user's library for the training phase, our model again beats the other models as Figures 5 and 6 show. As a matter of fact, many of the existing models actually work well under this setting, but poorly under the sparse setting. For example, CDL fails to beat POP in Citeulike-t dataset under the sparse setting, and then easily beats POP under the dense setting as Figures 3b and 5b show.

As a result, this experiment demonstrates the capability of our model to overcome the limitations mentioned in the beginning of this paper. For instance, among all the five

baseline models, CVAE++ has the best performance, which emphasizes the usefulness of involving more item's information, which are trained separately, to detect the latent factors of item more accurately. Also, the attentive autoencoder (AAE) can extract more constructive information over the variational autoencoder (VAE) and the stacked denoising autoencoder (SDAE) as CATA has the superiority over CVAE and CDL, and CATA++ has the superiority over CVAE++.

Table 5 shows the percentage of performance improvement of our model, CATA++, over the best competitor among all baselines. This percentage measures the increase in performance, which can be calculated according to the following formula: $improv\% = (p_{new} - p_{old})/p_{old} \times 100$, where p_{new} is the performance of our model, and p_{old} is the performance of the best model among all baselines.

In addition to the previous quantitative results, qualitative results to show the quality of recommendations using real examples are reported in Table 6. The table shows the top 10 recommendations of our model, CATA++, against the other competitive model, CVAE++, for one selected random user using Citeulike-2004-2007 dataset under the sparse setting. With this case study, we seek to gain a deeper insight into the difference between the two models in recommendations. The example in the table presents *user2214* who has only one article in his training library entitled "*A collaborative filtering framework based on fuzzy association rules and multiple-level similarity*". This example defines the sparsity problem very well where this user has limited feedback data. Based on the article's title, this user is probably interested in recommender systems and more specifically in collaborative

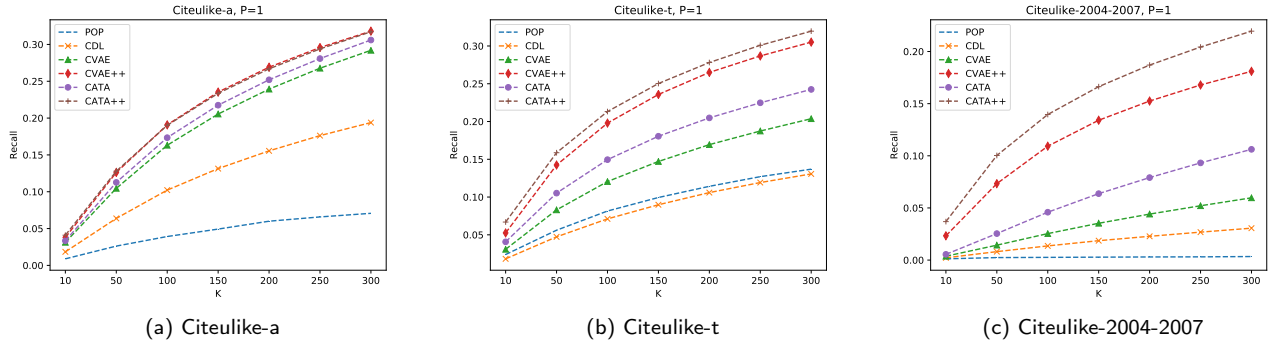


Figure 3: The top- K recommendation performance based on recall under the sparse setting, $P = 1$, for (a) Citeulike-a, (b) Citeulike-t, and (c) Citeulike-2004-2007 datasets.

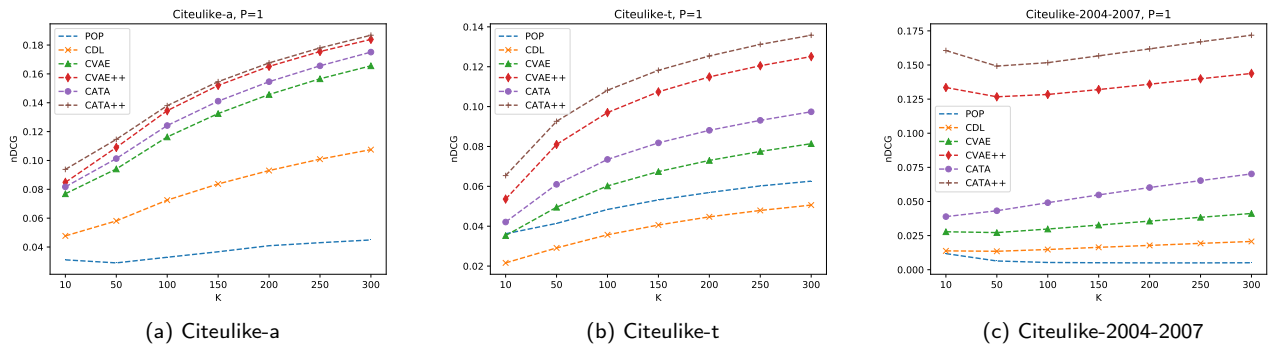


Figure 4: The top- K recommendation performance based on nDCG under the sparse setting, $P = 1$, for (a) Citeulike-a, (b) Citeulike-t, and (c) Citeulike-2004-2007 datasets.

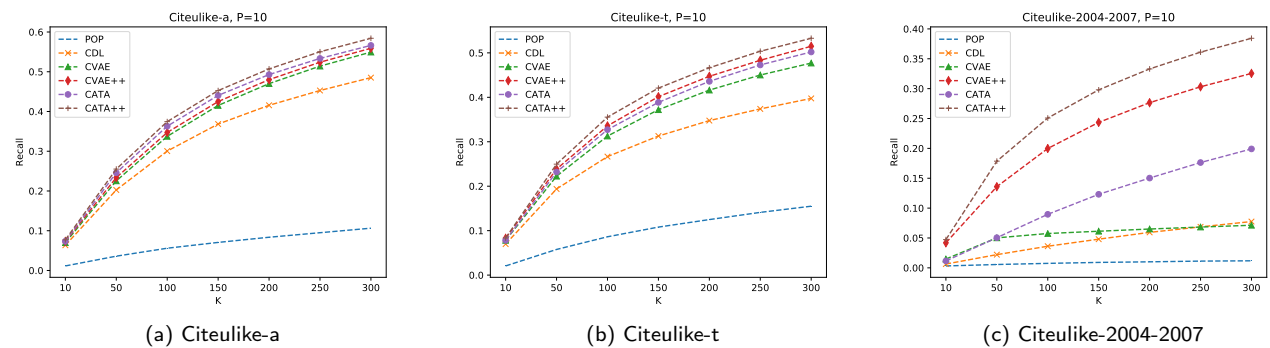


Figure 5: The top- K recommendation performance based on recall under the dense setting, $P = 10$, for (a) Citeulike-a, (b) Citeulike-t, and (c) Citeulike-2004-2007 datasets.

Table 5

Improvement percentage of the performance of CATA++ over the best competitor according to Recall@10, Recall@300, nDCG@10, and nDCG@300.

Approach	Sparse				Dense			
	Recall@10	Recall@300	nDCG@10	nDCG@300	Recall@10	Recall@300	nDCG@10	nDCG@300
Citeulike-a	8.18%	—	10.61%	1.57%	4.89%	3.16%	—	3.01%
Citeulike-t	27.42%	4.75%	22.01%	8.55%	0.84%	3.49%	—	3.15%
Citeulike-2004-2007	58.36%	21.27%	20.29%	19.47%	12.88%	18.06%	—	7.49%

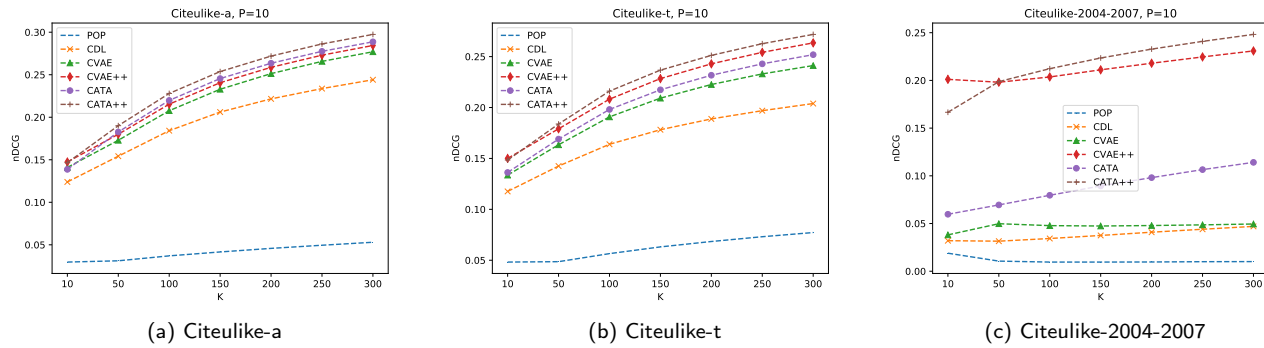


Figure 6: The top-K recommendation performance based on nDCG under the dense setting, $P = 10$, for (a) Citeulike-a, (b) Citeulike-t, and (c) Citeulike-2004-2007 datasets.

Table 6

A quality example of the top-10 recommendations under the sparse setting, $P = 1$, using Citeulike-2004-2007 dataset.

User ID: 2214	
Articles in training set: A collaborative filtering framework based on fuzzy association rules and multiple-level similarity	
CATA++	In user library?
1. Item-based collaborative filtering recommendation algorithms	No
2. Combining collaborative filtering with personal agents for better recommendations	No
3. An accurate and scalable collaborative recommender	No
4. Google news personalization: scalable online collaborative filtering	Yes
5. Combining collaborative and content-based filtering using conceptual graphs	Yes
6. Link prediction approach to collaborative filtering	No
7. Slope one predictors for online rating-based collaborative filtering	No
8. Slope one predictors for online rating-based collaborative filtering	Yes
9. A decentralized CF approach based on cooperative agents	No
10. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible..	Yes
CVAE++	In user library?
1. Combining collaborative filtering with personal agents for better recommendations	No
2. Explaining collaborative filtering recommendations	No
3. Google news personalization: scalable online collaborative filtering	Yes
4. Learning user interaction models for predicting web search result preferences	No
5. Item-based collaborative filtering recommendation algorithms	No
6. Enhancing digital libraries with TechLens+	No
7. Optimizing search engines using clickthrough data	No
8. Context-sensitive information retrieval using implicit feedback	No
9. A new approach for combining content-based and collaborative filters	No
10. Combining collaborative and content-based filtering using conceptual graphs	Yes

filtering (CF). After analyzing the results of each model, we can derive that our model can recommend more relevant articles than the other baseline. For instance, most of the top 10 recommendations based on CATA++ are related to the user's interest. The accuracy in this example is 0.4. Even though CVAE++ generates relevant articles as well, some irrelevant articles could be recommended as well such as the recommended article #7 entitled "*Optimizing search engines using clickthrough data*", which is more about search engines than RSs. From this example and other users' examples we have examined, we can state that our model detects the major elements of articles' contents and users' prefer-

ences more accurately, especially in the presence of limited data.

4.4.2. RQ2

To examine if the two autoencoders are cooperating with each other in finding more similarities between users and items, we run multiple experiments to show how each autoencoder performs solely compared to how they perform altogether. In other words, we compare the performance of using both autoencoders altogether in a parallel way (i.e., CATA++) against the performance of using only the right autoencoder (i.e., CATA) that leverages the articles' titles

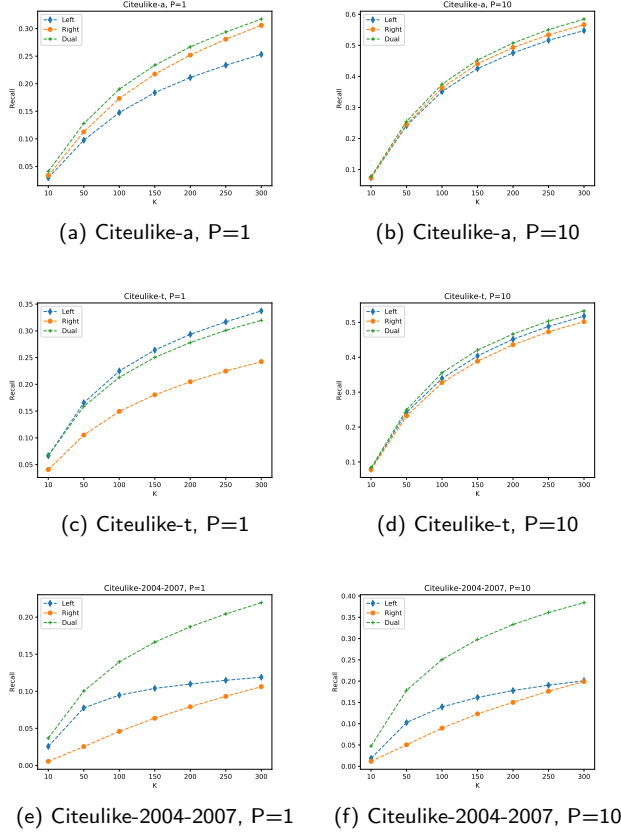


Figure 7: The performance results of using left autoencoder vs. right autoencoder compared to use both autoencoders altogether for (a-b) Citeulike-a, (c-d) Citeulike-t, and (e-f) Citeulike-2004-2007 datasets.

and abstracts, and against the performance of using only the left autoencoder that leverages the articles' tags and citations if available. Figure 7 shows the overall results. As the figure shows, the dual-way strategy has always better results than using each autoencoder solely except of one case in Figure 7c. In addition, the performance of the left autoencoder and the right autoencoder are competitive to each other such that the right autoencoder is better than the left autoencoder in Citeulike-a dataset, while the left autoencoder is better than the right autoencoder in the other two datasets. We can conclude that our model by coupling both autoencoders altogether is able to identify more accurate similarities between users and items which leads eventually to better recommendations.

4.4.3. RQ3

We conduct several experiments to find out the influence of tuning some hyper-parameters on the performance of our model, such as the dimension of the latent features, the number of hidden layers of the attentive autoencoder, and the two regularization parameters, λ_u and λ_v , used to learn the user/article latent features.

First, the dimension of the latent space used to report our results in the previous section is 50, i.e., each user and item

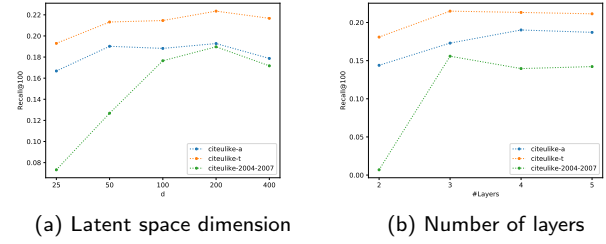


Figure 8: The impact of hyper-parameters tuning on CATA++ performance for: (a) dimension of features' latent space, and (b) number of layers inside each encoder and decoder.

latent feature, u_i and v_j , is a vector of size 50. We use the exact number as the state-of-the-art approach, CVAE, in order to have fair comparisons. However, to see the impact of different dimension sizes, we repeat our whole experiments by changing the size into one of following values {25, 50, 100, 200, 400}. In other words, we set the size of the latent factors of PMF and the size of the bottleneck of the attentive autoencoder to one of these values. As a result, we observe that when the dimension size is equal to 200, our model has the best performance on average among all three datasets as Figure 8a shows. Generally, setting the latent space with size between 100 and 200 is enough to have a reasonable performance compared to the other values.

Second, a four-layer network is used to construct our AAE when we report our results previously. The four-layer network has a shape of "#Vocabularies-400-200-100-50-100-200-400-#Vocabularies". However, we again repeat the whole experiments with different number of layers starting from 2 to 5 layers, such that each layer has a half size of the previous one. As Figure 8b shows, using less than 3 layers are not enough to learn the side information. Generally, 3-layer and 4-layer networks are good enough to train our model.

Third, we repeat the experiment again with different values of λ_u and λ_v from the following range {0.01, 0.1, 1, 10, 100}. Figures 9a and 9c show the performance under the sparse setting in Citeulike-a and Citeulike-t datasets respectively. From these two figures, using a lower value of λ_v typically results into lower performance. That means the user feedback data is not enough and it needs more article information. Same thing can be said to both scenarios of Citeulike-2004-2007 dataset in Figures 9e and 9f. Additionally in Figure 9e, higher value of λ_u decrease the performance where user feedback is scarce. Even though Figure 9f shows the performance under the dense setting for Citeulike-2004-2007 dataset, it still exemplifies the sparsity with regard to articles as we indicate before in Figure 2, where 80% of the articles have been only added to 1 user library. On the other hand where user feedback is considerably enough, higher value of λ_v results into lower performance as Figures 9b and 9d show.

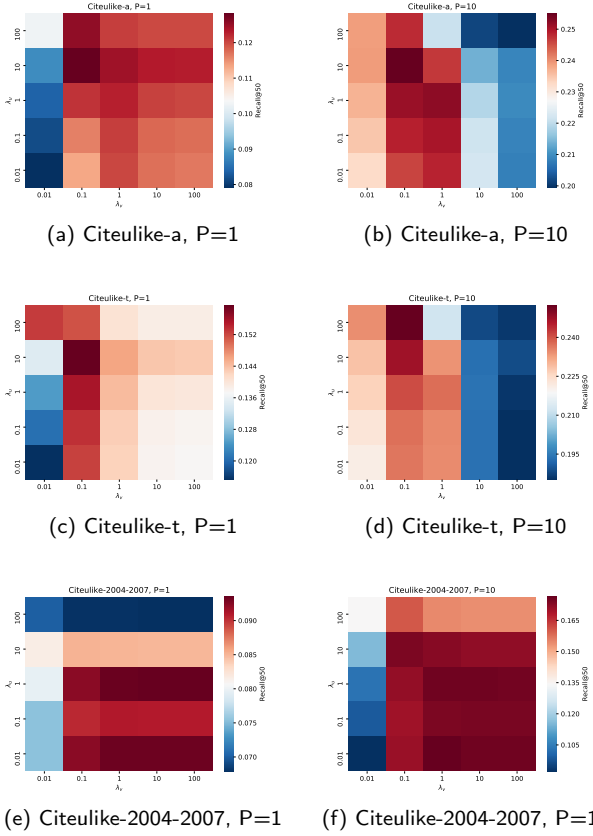


Figure 9: The impact of λ_u and λ_v on CATA++ performance for (a-b) Citeulike-a, (c-d) Citeulike-t, and (e-f) Citeulike-2004-2007 datasets.

5. Conclusion

In this paper, we alleviate the natural data sparsity problem in recommender systems by introducing a dual-way strategy to learn item's textual information by coupling two parallel attentive autoencoders together. The learned item's features are then utilized in the learning process of matrix factorization (MF). We evaluate our model for academic article recommendation task using three real-world datasets. The huge gap in the experimental results validates the usefulness of exploiting more item's information, and the benefit of integrating attention technique in finding more relevant recommendations, and thus boosting the recommendation accuracy. As a result, our model, CATA++, has the superiority over multiple state-of-the-art MF based models according to various evaluation metrics. Furthermore, the performance improvement of CATA++ increases consistently as the data sparsity increases from one dataset to another.

For future work, new metric learning algorithms could be explored to substitute MF technique because the dot product in MF doesn't guarantee the triangle inequality [6]. For any three items, the triangle inequality is fulfilled once the sum of distance between any two item pairs in the feature space should be greater or equal to the distance of the third item pair, such that $d(x, y) \leq d(x, z) + d(z, y)$. By doing

so, user-user and item-item relationships might be captured more accurately.

CRediT authorship contribution statement

Meshal Alfarhood: Conceptualization, Methodology, Software, Validation, Writing - original draft. **Jianlin Cheng:** Conceptualization, Supervision, Resources, Writing - review & editing.

References

- [1] Alfarhood, M., Cheng, J., 2019. Collaborative attentive autoencoder for scientific article recommendation, in: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE. pp. 168–174.
- [2] Alzogbi, A., 2018. Time-aware collaborative topic regression: Towards higher relevance in textual item recommendation., in: BIRNDL@ SIGIR, pp. 10–23.
- [3] Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 .
- [4] He, X., He, Z., Song, J., Liu, Z., Jiang, Y., Chua, T., 2018. Nais: Neural attentive item similarity model for recommendation. IEEE Transactions on Knowledge and Data Engineering 30, 2354–2366.
- [5] Hinton, G.E., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. science 313, 504–507.
- [6] Hsieh, C., Yang, L., Cui, Y., Lin, T., Belongie, S., Estrin, D., 2017. Collaborative metric learning, in: Proceedings of the 26th international conference on world wide web, International World Wide Web Conferences Steering Committee. pp. 193–201.
- [7] Hu, Y., Koren, Y., Volinsky, C., 2008. Collaborative filtering for implicit feedback datasets., in: ICDM, Citeseer. pp. 263–272.
- [8] Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .
- [9] Jhamb, Y., Ebesu, T., Fang, Y., 2018. Attentive contextual denoising autoencoder for recommendation, in: Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ACM. pp. 27–34.
- [10] Jing, L., Wang, P., Yang, L., 2015. Sparse probabilistic matrix factorization by laplace distribution for collaborative filtering, in: Twenty-Fourth International Joint Conference on Artificial Intelligence.
- [11] Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. Computer , 30–37.
- [12] Li, S., Kawale, J., Fu, Y., 2015. Deep collaborative filtering via marginalized denoising auto-encoder, in: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, ACM. pp. 811–820.
- [13] Li, X., She, J., 2017. Collaborative variational autoencoder for recommender systems, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 305–314.
- [14] Ma, C., Kang, P., Wu, B., Wang, Q., Liu, X., 2019. Gated attentive-autoencoder for content-aware recommendation, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, ACM. pp. 519–527.
- [15] 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.
- [16] Mnih, A., Salakhutdinov, R., 2008. Probabilistic matrix factorization, in: Advances in neural information processing systems, pp. 1257–1264.
- [17] Mnih, V., Heess, N., Graves, A., Kavukcuoglu, K., 2014. Recurrent models of visual attention, in: Advances in neural information processing systems, pp. 2204–2212.

- [18] Pan, R., Zhou, Y., Cao, B., Liu, N.N., Lukose, R., Scholz, M., Yang, Q., 2008. One-class collaborative filtering, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE. pp. 502–511.
- [19] 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.
- [20] Salakhutdinov, R., Mnih, A., 2008. Bayesian probabilistic matrix factorization using markov chain monte carlo, in: Proceedings of the 25th international conference on Machine learning, ACM. pp. 880–887.
- [21] Salakhutdinov, R., Mnih, A., Hinton, G., 2007. Restricted boltzmann machines for collaborative filtering, in: Proceedings of the 24th international conference on Machine learning, ACM. pp. 791–798.
- [22] Salton, G., McGill, M., 1983. Introduction to modern information retrieval. McGraw-Hill, Inc.
- [23] 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.
- [24] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929–1958.
- [25] Tay, Y., Luu, A., Hui, S., 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.
- [26] Wang, C., Blei, D., 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.
- [27] Wang, H., Chen, B., Li, W., 2013. Collaborative topic regression with social regularization for tag recommendation, in: Twenty-Third International Joint Conference on Artificial Intelligence.
- [28] Wang, H., Wang, N., Yeung, D., 2015. Collaborative deep learning for recommender systems, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. pp. 1235–1244.
- [29] Xiao, J., Ye, H., He, X., Zhang, H., Wu, F., Chua, T., 2017. Attentional factorization machines: learning the weight of feature interactions via attention networks, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press. pp. 3119–3125.