Gated Attentive-Autoencoder for Content-Aware Recommendation

Chen Ma¹, Peng Kang¹, Bin Wu², Qinglong Wang¹ and Xue Liu¹

¹McGill University, Montreal, Canada
²Zhengzhou University, Zhengzhou, China

WSDM2019, Melbourne, Australia
Background

The rapid growth of Internet services allows users to access millions of online products, such as movies, articles.

The large amount of user-item data facilitates a promising and practical service – the **personalized recommendation**.
The rapid growth of Internet services allows users to access millions of online products, such as movies, articles.

<table>
<thead>
<tr>
<th>movie</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
The rapid growth of Internet services allows users to access millions of online products, such as movies, articles.

<table>
<thead>
<tr>
<th>movie</th>
<th>user 1</th>
<th>user 2</th>
<th>user 3</th>
<th>user 4</th>
<th>user 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>2</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>4</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>5</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Background

The rapid growth of Internet services allows users to access millions of online products, such as movies, articles.

<table>
<thead>
<tr>
<th>user</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

- I
- II
- III
Background

The rapid growth of Internet services allows users to access millions of online products, such as movies, articles.

<table>
<thead>
<tr>
<th>movie</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Content helps
Less privacy issue
Related Work

<table>
<thead>
<tr>
<th>Models</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTR (Wang et al., SIGKDD’ 2011)</td>
<td>MF + LDA</td>
</tr>
<tr>
<td>SVDFeature (Chen et al., JMLR’ 2012)</td>
<td>Feature-based MF</td>
</tr>
<tr>
<td>HFT (Julian et al., RecSys’ 2013)</td>
<td>LFM + LDA</td>
</tr>
<tr>
<td>CDL (Wang et al., SIGKDD’ 2015)</td>
<td>MF + SDAE</td>
</tr>
<tr>
<td>ConvMF (Kim et al., RecSys’ 2016)</td>
<td>MF + CNN</td>
</tr>
<tr>
<td>CVAE (Li et al., SIGKDD’ 2017)</td>
<td>MF + VAE</td>
</tr>
</tbody>
</table>

- Equally treat item content
- Combine the rating and content information by regularization
- Not explicitly utilize the item-item relations

MF: Matrix Factorization  
LDA: Latent Dirichlet Allocation  
LFM: Latent Factor Model  
SDAE: Stacked Denoising AutoEncoder  
VAE: Variational AutoEncoder
Model Overview

An autoencoder-based model:

Word-attention

Gating layer

Neighbor-attention
Model Overview

An autoencoder-based model:

Word-attention

Gating layer

Neighbor-attention
Autoencoder

- Autoencoder is used to learn the item hidden representations from rating information.

\[
\begin{align*}
\text{enc} : & \quad \begin{cases} 
    z_i^{(1)} = a_1(W_1 r_i + b_1) \\
    z_i^r = a_2(W_2 z_i^{(1)} + b_2)
\end{cases} \\
\text{dec} : & \quad \begin{cases} 
    z_i^{(3)} = a_3(W_3 z_i^r + b_3) \\
    \hat{r}_i = a_4(W_4 z_i^{(3)} + b_4)
\end{cases} \\
\text{loss} : & \quad \mathcal{L} = \sum_{i=1}^{M} \| x_i - \hat{x}_i \|_2^2
\end{align*}
\]
Word-attention Module

Binary Item Ratings

Encoder_r

Gating Layer

G
1 - G

One-hop neighbors

Neighbor_Att

Decoder_r

Word Embeddings $D_i$

Item Content
The Avengers are amazing...

Word_Att

Agg_Layer

$z_i^r$  1 1 0 0 0 0 0

$r_i$

$\hat{r}_i$ 0.8 0.9 0.2 0.1 0.1 0.1 0.2
Word-attention Module

• Previous works do not discriminate the word importance for describing a certain item
• Some informative words are more representative than others and should contribute more to characterize a certain item

E.G. great food and good service .... what else can you ask for

• We utilize an attention model to learn the item representation from content information.
Word-attention Module

Item Content

The Avengers are amazing ...

\[ D_i = \begin{bmatrix} \ldots & e_{j-1} & e_j & e_{j+1} & \ldots \end{bmatrix} \]

word embedding look-up

attention score matrix

\[ A_i = softmax(W_{a_1} \tanh(W_{a_2} D_i + b_{a_2}) + b_{a_1}) \]

matrix representation of items

\[ Z_i^c = A_i D_i^T \]

aggregate item representations into one aspect

\[ z_i^c = a_t(Z_i^c^T w_t) \]
Gating Layer

Binary Item Ratings

Encoder_r

Gating Layer

Decoder_r

One-hop neighbors

Neighbor_Att

Word Embeddings $D_i$

Item Content

The Avengers are amazing ...

Word_Att

Agg_Layer

$r_i$ 1 1 0 0 0 0 0

$z_i^r$

$\hat{r}_i$ 0.8 0.9 0.2 0.1 0.1 0.1 0.2
Gating Layer

• Adaptively fuse the hidden representations from two heterogeneous data sources

• Avoid tedious hyper-parameter tuning by the regularization term

\[ G = \text{sigmoid}(W_{g1}z_i^r + W_{g2}z_i^c + b_g) \] adaptively learn the gate

\[ z_i^g = G \odot z_i^r + (1 - G) \odot z_i^c \] combine hidden representations
Neighbor-attention Module

Binary Item Ratings

Encoder_r

Gating Layer

Gating

1 - G

One-hop neighbors

Neighbor_Att

Decoder_r

\( \tilde{r}_i \)

0.8 0.9 0.2 0.1 0.1 0.1 0.2

Item Content

The Avengers are amazing...

Word Embeddings

D_i

e_1 e_2 e_3 e_4 ...

Word_Att

Agg_Layer
Neighbor-attention Module

• Previous works do not consider the relations between items
• Related items may share the same topic or have similar attributes: citations between articles, movies in the same genre
• Exploring users’ preferences on an item’s neighbors also benefits inferring users’ preferences on this item
Neighbor-attention Module

\[ s_{i,j} = \tanh(z_i^g W_n z_j^g), \forall j \in N_i \]  
use a bilinear function to capture the relation

\[ a_i = \text{softmax}(s_i) \]  
the attention score of item i’s neighbors

\[ z_i^n = \sum_{j \in N_i} a_{i,j} z_j^g \]  
the neighborhood representation item i
Prediction and Loss

• Modified decoder: explore users’ preferences on both an item and its neighborhood

\[
\begin{align*}
  z_i^{(3,g)} &= a_3(W_3z_i^g + b_3), \\
  z_i^{(3,n)} &= a_3(W_3z_i^n + b_3), \\
  \hat{r}_i &= a_4(W_4z_i^{(3,g)} + W_4z_i^{(3,n)} + b_4).
\end{align*}
\]

• Weighted loss

\[
\mathcal{L}_{AE} = \sum_{i=1}^{n} \sum_{u=1}^{m} \left\| C_{u,i}(R_{u,i} - \hat{R}_{u,i}) \right\|_2^2 = \left\| C^T \odot (R^T - \hat{R}^T) \right\|_F^2 \quad C_{u,i} = \begin{cases} 
  \rho & \text{if } R_{u,i} = 1 \\
  1 & \text{otherwise}
\end{cases}
\]
Evaluation

• Four datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Users</th>
<th>#Items</th>
<th>#Ratings</th>
<th>#Words</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>citeulike-a</td>
<td>5,551</td>
<td>16,980</td>
<td>204,986</td>
<td>8,000</td>
<td>0.217%</td>
</tr>
<tr>
<td>ML20M</td>
<td>138,493</td>
<td>18,307</td>
<td>19,977,049</td>
<td>12,397</td>
<td>0.788%</td>
</tr>
<tr>
<td>Books</td>
<td>65,476</td>
<td>41,264</td>
<td>1,947,765</td>
<td>27,584</td>
<td>0.072%</td>
</tr>
<tr>
<td>CDs</td>
<td>24,934</td>
<td>24,634</td>
<td>478,048</td>
<td>24,341</td>
<td>0.078%</td>
</tr>
</tbody>
</table>

For each user, 20% of her viewed items are selected as testing.

• Evaluation Metrics
  • Recall@5, 10, 15, 20
  • NDCG @5, 10, 15, 20
Evaluation Baselines

WRMF: weighted regularized matrix factorization, ICDM’ 2008
CDAE: collaborative denoising autoencoder, WSDM’ 2016
CDL: collaborative deep learning, SIGKDD’ 2015
CVAE: collaborative variational autoencoder, SIGKDD’ 2015
CML+F: collaborative metric learning, WWW’ 2017
ConvMF: convolutional matrix factorization, RecSys’ 2016
JRL: joint representation learning, CIKM’ 2017
### Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>WRMF</th>
<th>CDAE</th>
<th>CDL</th>
<th>CVAE</th>
<th>CML+F</th>
<th>ConvMF</th>
<th>JRL</th>
<th>GATE</th>
<th>Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recall@10</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>citeulike-a</td>
<td>0.0946</td>
<td>0.0888</td>
<td>0.1317</td>
<td>0.1371</td>
<td>0.1283</td>
<td>0.1153</td>
<td>0.1325</td>
<td><strong>0.1419</strong></td>
<td>3.50%</td>
</tr>
<tr>
<td>movielens-20M</td>
<td>0.1075</td>
<td>0.0751</td>
<td>0.1287</td>
<td>0.1303</td>
<td>0.1123</td>
<td>0.1201</td>
<td>0.1401</td>
<td><strong>0.1625</strong></td>
<td><strong>15.99%</strong></td>
</tr>
<tr>
<td>Amazon-Books</td>
<td>0.0553</td>
<td>0.0132</td>
<td>0.0648</td>
<td>0.0632</td>
<td>0.0756</td>
<td>0.0524</td>
<td><strong>0.0924</strong></td>
<td><strong>0.1133</strong></td>
<td><strong>22.62%</strong></td>
</tr>
<tr>
<td>Amazon-CDs</td>
<td>0.0779</td>
<td>0.0191</td>
<td><strong>0.0827</strong></td>
<td>0.0811</td>
<td>0.0824</td>
<td>0.0753</td>
<td>0.0816</td>
<td><strong>0.1057</strong></td>
<td><strong>27.81%</strong></td>
</tr>
</tbody>
</table>

|                  |        |        |        |        |        |        |        |        |         |
| **NDCG@10**      |        |        |        |        |        |        |        |        |         |
| citeulike-a      | 0.0843 | 0.0736 | 0.0949 | 0.0952 | **0.1035** | 0.0914 | 0.0982 | **0.1082** | **4.54%** |
| movielens-20M    | 0.1806 | 0.1774 | 0.1836 | 0.1939 | **0.2479** | 0.1807 | 0.2439 | **0.2992** | **20.69%** |
| Amazon-Books     | 0.0377 | 0.0105 | 0.0393 | 0.0384 | 0.0456 | 0.0324 | **0.0592** | **0.0708** | **19.59%** |
| Amazon-CDs       | 0.0357 | 0.0105 | 0.0356 | 0.0349 | 0.0364 | 0.0323 | **0.0386** | **0.0477** | **23.58%** |

*: p <= 0.05, **p < 0.01, ***: p < 0.001

Our GATE outperforms other methods significantly on most of the datasets.
Evaluation Results

• Ablation study

<table>
<thead>
<tr>
<th>Architecture</th>
<th>CDs</th>
<th>Books</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@10</td>
<td>N@10</td>
</tr>
<tr>
<td>(1) stacked AE</td>
<td>0.0672</td>
<td>0.0315</td>
</tr>
<tr>
<td>(2) reg: AE + W_Att</td>
<td>0.0676</td>
<td>0.0318</td>
</tr>
<tr>
<td>(3) gating: AE + W_Att</td>
<td>0.0816</td>
<td>0.0353</td>
</tr>
<tr>
<td>(4) gating: AE + GRU</td>
<td>0.0818</td>
<td>0.0352</td>
</tr>
<tr>
<td>(5) gating: AE + CNN</td>
<td>0.0777</td>
<td>0.0335</td>
</tr>
<tr>
<td>(6) GATE</td>
<td><strong>0.1057</strong></td>
<td><strong>0.0477</strong></td>
</tr>
</tbody>
</table>

From (2), (3): our gating is better than regularization
From (3), (4), (5): our word-attention achieves similar performance with fewer parameters
From (3), (6): the item-item relations play an important role
Evaluation Results

• Case Study

<table>
<thead>
<tr>
<th>Target</th>
<th>Neighbor</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Genomic analysis of regulatory network dynamics reveals large topological changes (0)</td>
<td>0.07172</td>
</tr>
<tr>
<td></td>
<td>Frequency of occurrence of numbers in the World Wide Web (10)</td>
<td>0.22090</td>
</tr>
<tr>
<td></td>
<td>Complex networks: Structure and dynamics (16)</td>
<td>0.26835</td>
</tr>
<tr>
<td></td>
<td>Noise in protein expression scales with natural protein abundance (36)</td>
<td><strong>0.43903</strong></td>
</tr>
</tbody>
</table>

The Summary of Article 16797 in citeulike-a

We present the first parallel implementation of the T-Coffee consistency-based multiple aligner. We benchmark it on the Amazon Elastic Cloud (EC2) and show that the parallelization procedure is reasonably effective. We also conclude that for a web server with moderate usage (10K hits/month) the cloud provides a cost-effective alternative to in-house deployment.
Conclusion

We propose an autoencoder-based model, which consists of a word-attention module, a neighbor-attention module, and a gating layer to address the content-aware recommendation task.

Experimental results show that the proposed method outperforms the state-of-the-art methods significantly for content-aware recommendation.
Thank you!

Q & A

Email: chen.ma2@mail.mcgill.ca
Code: https://github.com/allenjack/GATE
LibRec: https://www.librec.net/

Google ‘LibRec’