

Adversarial Distillation for Efficient Recommendation with External Knowledge

XU CHEN, School of Software, Tsinghua University

YONGFENG ZHANG, Department of Computer Science, Rutgers University

HONGTENG XU, Department of ECE, Duke University

ZHENG QIN, School of Software, Tsinghua University

HONGYUAN ZHA, College of Computing, Georgia Institute of Technology

ACM Transactions on Information Systems 2018

Zhang Junwei

2019/6/28₁

INTRODUCTION

- External knowledge
- in this article, we would like to ask: *“Can we not only take the power of deep learning for external knowledge modeling, but also keep a runtime-efficient model?”*

To answer this question, we reformulate the problem of recommendation with external knowledge into a **generalized distillation framework** (GDF), where the **complex architecture** (e.g., CNN) is moved into a separate component, which is only used in the training phrase, while abandoned at test time.

By connecting the top layers between the teacher and student models, the valuable information learned by the teacher model is transfered (distilled) into the student model to help enhance its prediction ability.

Table 1. A High-Level Comparison Between Our Model with Previous Methods

Properties	HFT	RBLT	CTR	DeepCoNN	TransNet	SDNet
Reference	[31]	[41]	[43]	[51]	[3]	-
Model Depth	shallow	shallow	shallow	deep	deep	deep
Effectiveness	↓	↓	↓	↑	↑	↑
Runtime Efficiency	↑	↑	↑	↓	↓	↑

↑ means relative high and ↓ means relative low.

As an implementation of this framework, we select **user review as the external knowledge**. The overall principle can be seen in Figure 1(b). The student model is specified as a user-item prediction network, which **takes a pair of user-item IDs as input and output a rating**, while the teacher model is implemented by a review prediction network, which maps a piece of user review into a rating.

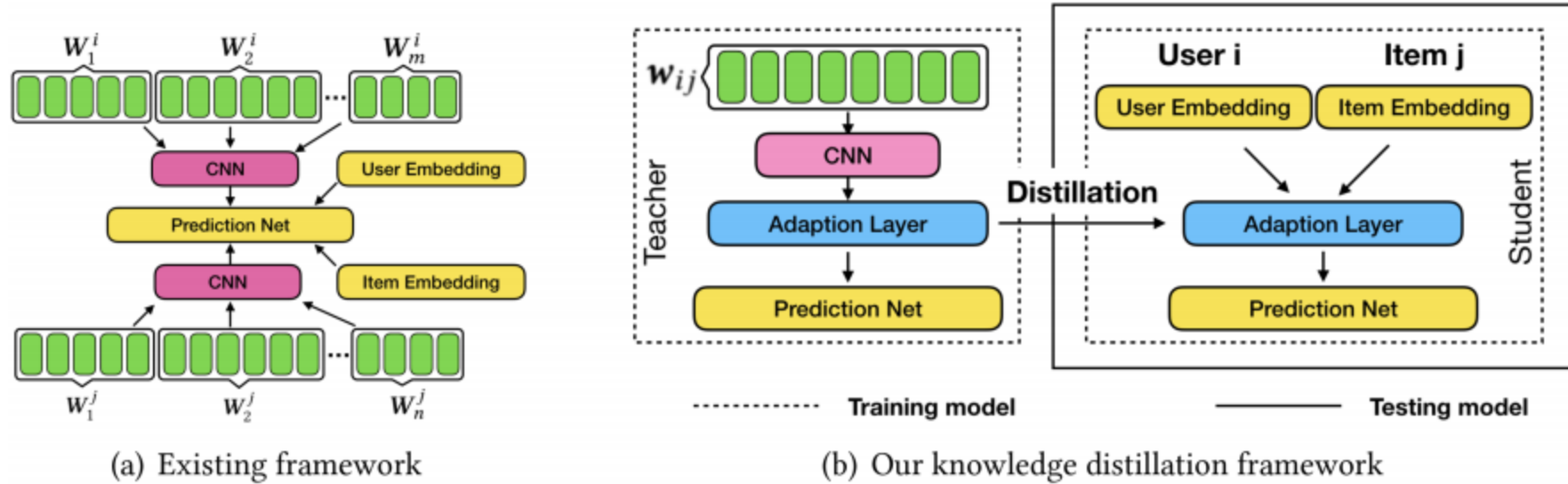
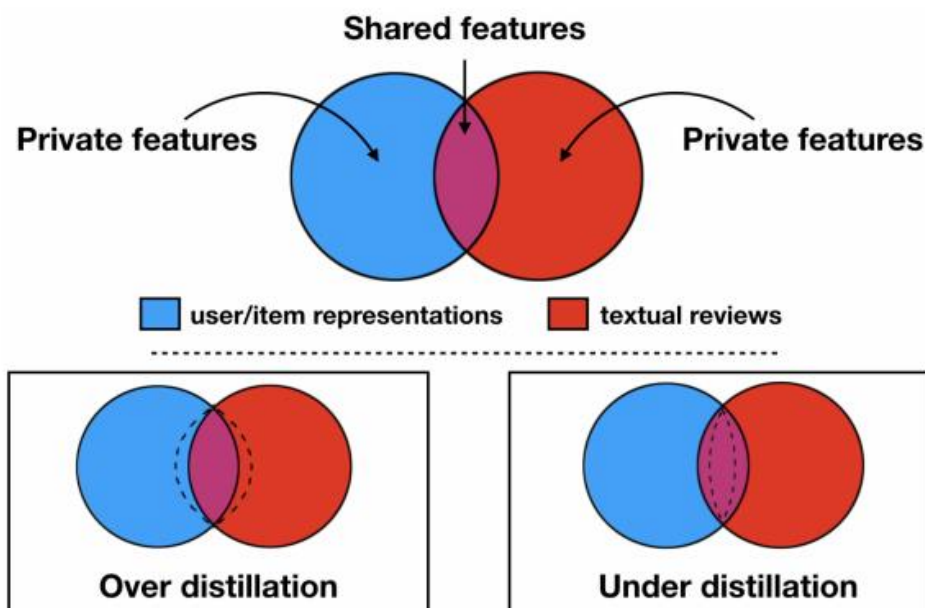



Fig. 1. (a) The core architecture of many existing review-based recommendation models. $\{W_1^i, W_2^i, \dots, W_m^i\}$ and $\{W_1^j, W_2^j, \dots, W_n^j\}$ are the reviews of user i and item j in the training set, which are respectively concatenated before inputting into the CNN networks to make the final rating prediction. (b) Our idea of decomposing user review modeling into a teacher model, which only exists in the training phase, and will not be leveraged when making predictions. w_{ij} is the review from user i to item j .


In the task of review-based recommendation, such methods can be **less effective** because the review information processed by the teacher model and the user/item properties encoded in the student model do not always agree with each other: on one hand, many words in user reviews are **not directly related to user preferences** (or item features); On the other hand, many user preferences (or item features) **cannot be discovered from review information**



Product Images



onlypuff Casual Shirts For Women
Tunic Tops Pullover Sweatshirts
Blouse Solid Color Bottom Basic Tops
★★★★☆ · 165 customer reviews
| 8 answered questions



MOSPRO Trail Camera Viewer for
iPhone iPad Mac & Android, SD &
Micro SD Memory Card Reader to
View Photos and Videos from any
Wildlife Scouting Game Cam on
Smartphone for Deer Hunter Black
★★★★★ · 106 customer reviews
| 36 answered questions

★★★★☆ was very disappointed. I gave it away to my niece
By [Kindle Customer](#) on November 30, 2017
Size: XXXX-Large | Color: Orange | **Verified Purchase**

I ordered it for my birthday. The sleeves were too small, was very disappointed. I gave it away to my niece.

★★★★★ Works great. Hunting hubby very pleased!
December 31, 2017
Color: Yellow | **Verified Purchase**

I had bought this for my husband so he could easily check his game camera while waiting in the tree stand. The application download was very simple. Picture quality is great. Very simple to use. I would recommend buying the case that goes with this for protection of the device while in the woods.

User Reviews

To transfer just the right amount of valuable knowledge from the review information into the user/item representations, we carefully design a novel Selective Distillation Network (SDNet for short) tailored for the task of review-based recommendation.

In particular, the top layer in each of the teacher and student models is designed to contain **two types of features**: one is the **shared features**, which are used for transferring valuable review information. The other is the **private features**, which are reserved to capture the mismatching knowledge.

Ideally, the shared features should only transfer valuable knowledge, and at the same time, all the valuable knowledge should be included in the shared features.

we design the adversarial adaption strategy to connect the teacher and student models for effective knowledge transfer. On the other hand, we adopt the orthogonality constraint to forcefully have the shared and private features in each component encode different information, so as to push the valuable knowledge leaked in the private features into the shared features following the supervision signal.

Contributions

We propose to reformulate recommendation with external knowledge into a generalized distillation framework, based on which we can not only take advantage of deep architectures for external knowledge modeling, but also can keep the proposed model computationally efficient at test time.

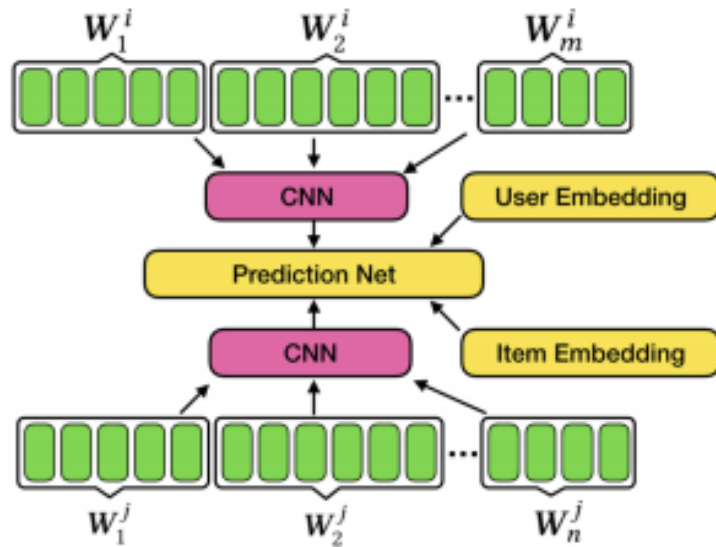
- We design a novel neural network, SDNet, for [solving the mismatching problem between the teacher and student model](#) in the generalized distillation framework. And we apply this method to the task of review-based recommendation for performance evaluation.
- We conduct extensive experiments to verify that, compared with the state-of-the-art methods, our models are not only able to improve the recommendation performance, but also can greatly reduce the computational time when making predictions.

PRELIMINARIES

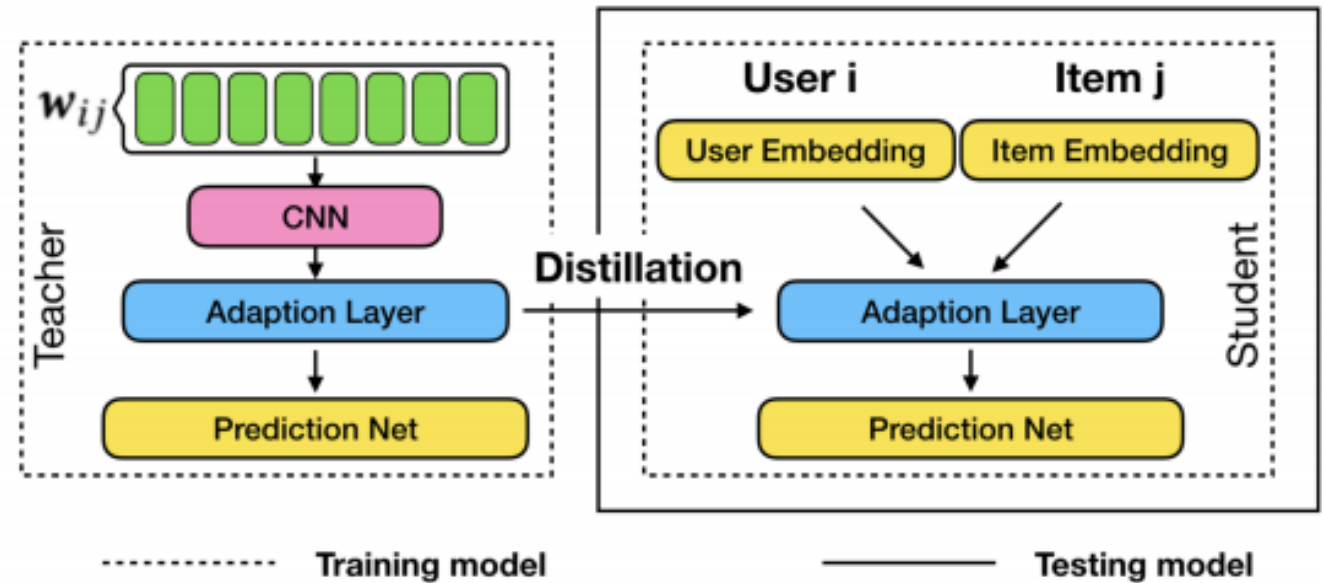
● Review-Based Recommendation

Notations	Descriptions
\mathcal{U}	The set of N users $\{u_1, u_2, \dots, u_N\}$.
\mathcal{V}	The set of M items $\{v_1, v_2, \dots, v_M\}$.
D	The embedding dimension.
$\tilde{\mathbf{p}}_i, \bar{\mathbf{p}}_i$	The private and shared embedding of user i .
$\tilde{\mathbf{q}}_j, \bar{\mathbf{q}}_j$	The private and shared embedding of item j .
r_{ij}, \mathcal{R}	The rating scored by user i on item j , and the set of all ratings.
$l_{ij}, \mathbf{w}_{ij}, \mathcal{W}$	The length and word list of user i 's review on item j : $\{w_{ij}^1, w_{ij}^2, \dots, w_{ij}^{l_{ij}}\}$, and the set of all reviews $\{\mathbf{w}_{ij} i \in \mathcal{U}, j \in \mathcal{V}\}$.
$\mathbf{W}^i, \mathbf{W}^j$	The set of reviews related to user i and item j .
λ	The regularization coefficient in the user-item prediction network.
b_0, b_i, \mathbf{a}_i	The parameters in the FM layer of user-item prediction network.
$\mathbf{t}_{ij}, \mathbf{s}_{ij}$	The complete private and shared embedding.
$\mathbf{t}_{w_{ij}}, \mathbf{s}_{w_{ij}}$	The outputs from shared and private review processors.
\mathbf{V}, g	The weighting and bias parameters in the review prediction network.
$\mathbf{V}^i, \mathbf{V}^f, \mathbf{V}^o, \mathbf{V}^c$	The weighting parameters in LSTM.
$\mathbf{g}^i, \mathbf{g}^f, \mathbf{g}^o, \mathbf{g}^c$	The bias parameters in LSTM.
t, f_j	The window size, and the output from the j -th filter in CNN.
λ_{1-4}	The weighting parameters that balance different objective functions.
σ, \tanh, h	The sigmoid, hyperbolic tangent, and ReLU activation functions.

Based on these analyses, we regard user review as privileged information, and revisited the task of review-based recommendation by a generalized distillation framework.



(a) Existing framework



(b) Our knowledge distillation framework

- **Generalized Distillation Framework**

$$r_{ij} = g(\text{CNN}(\text{concat}(\mathbf{W}^i)), \text{CNN}(\text{concat}(\mathbf{W}^j))),$$

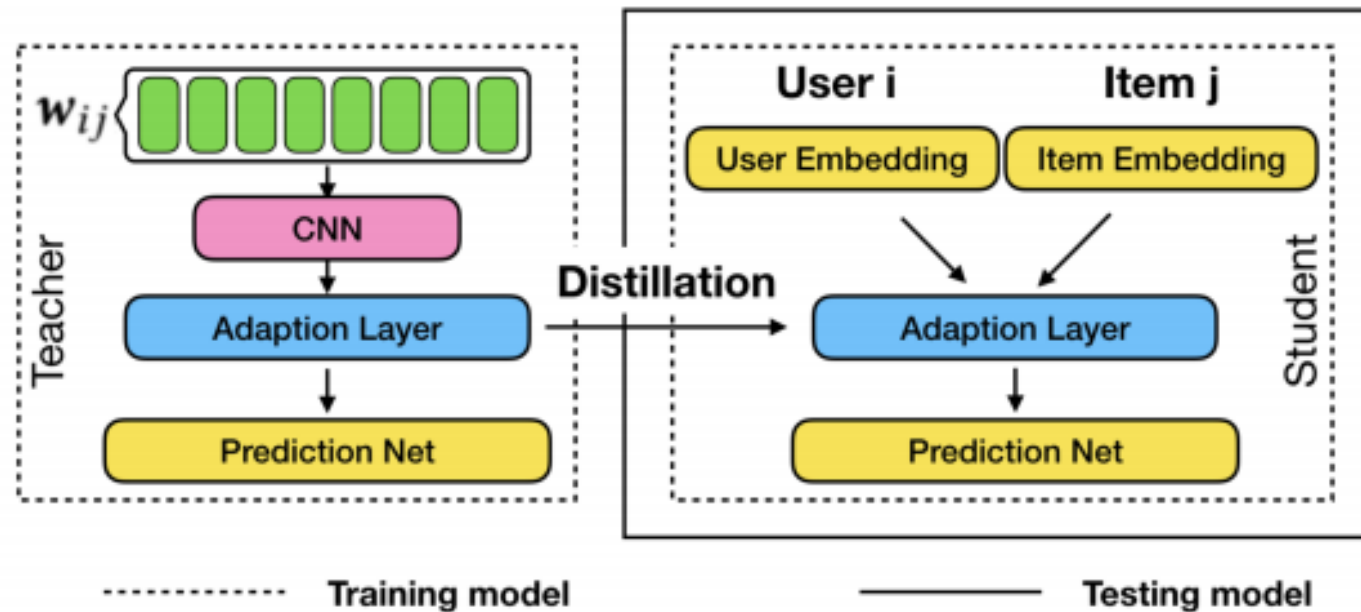
$$h_T = \arg \min_{h \in \mathcal{H}_T} \frac{1}{n} \sum_{t=1}^n \ell(y_t, h(x_t^*)) + \Omega(\|h\|),$$

$$h_S = \arg \min_{h \in \mathcal{H}_S} \frac{1}{n} \sum_{t=1}^n [\lambda \ell(y_t, h(x_t)) + (1 - \lambda) \ell(h_T(x_t^*), h(x_t))],$$

● Review-Based Recommendation as Generalized Distillation Framework

we re-organize the user behavioral data in the above settings as a set of quadruples $O = \{(i, j, w_{ij}, r_{ij})_t\}_{t=1}^n$, where the element (i, j, w_{ij}, r_{ij}) means user i interacts with item j by reviewing w_{ij} and rating r_{ij} . The rationalities of formulating review-based recommendation as a generalized distillation framework comes from two aspects:

- Under the generalized distillation framework, the complex model component can be separated into the teacher model, which is only used in the training phrase, and does not influence the runtime efficiency.
- In a review-based recommender system, the ID information (e.g., i, j) is mainly used to distinguish different users (or items), and its rating prediction ability comes from the collaborative filtering assumption. Whereas, the review information (e.g., w_{ij}) can explicitly reflect more comprehensive user/item properties and reveal the specific reasons for the predicted results (e.g., r_{ij}).



(b) Our knowledge distillation framework

Although this idea seems promising, user reviews are usually very noisy in real settings, and directly transferring (distilling) all the textual information into the student model can be less effective.

- **adversarial adaption**
- **and orthogonality constraint strategies**

SDNET: SELECTIVE DISTILLATION NETWORK

● Model Structure

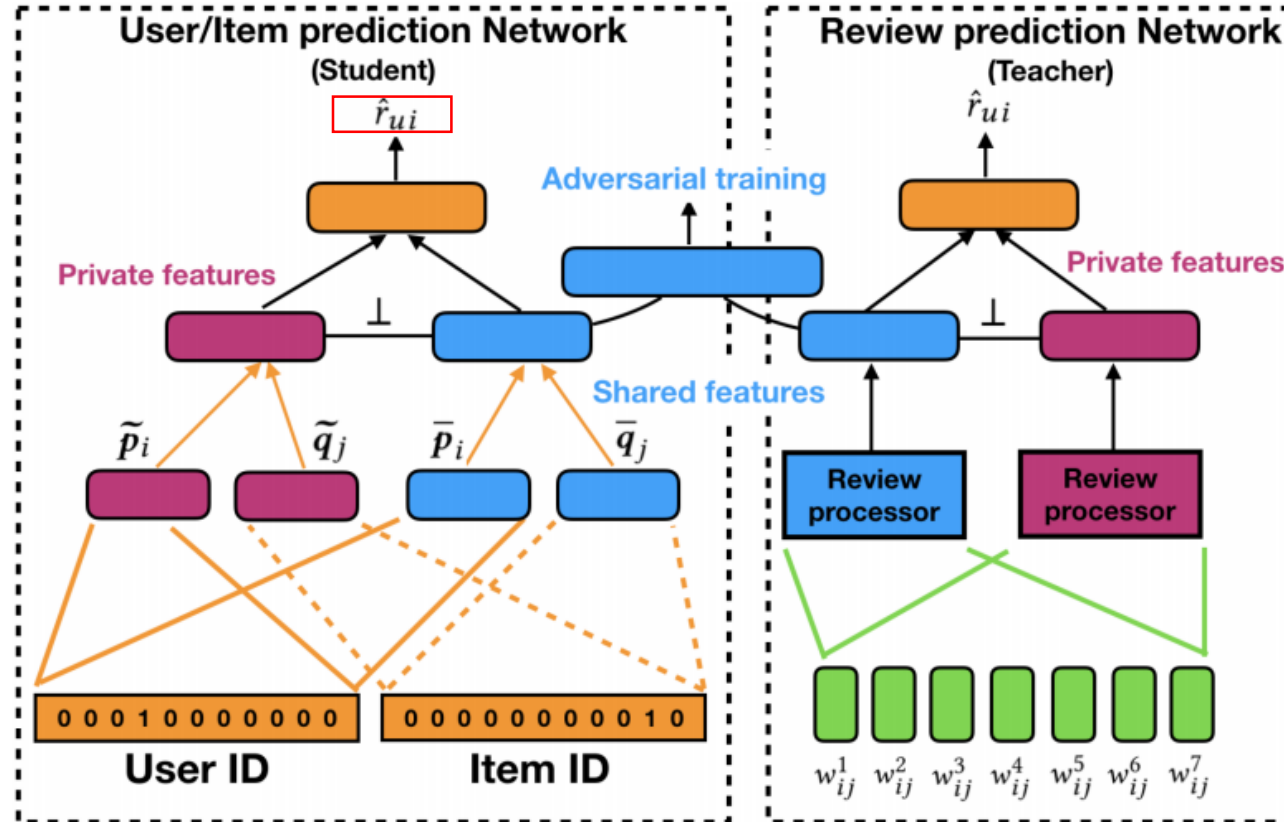


Fig. 4. Our overall framework. The left user-item prediction network and the right review prediction network are connected in an adversarial manner, and the shared and private features in the top layer of each network are encouraged to be orthogonal (denoted as \perp).

User-Item Prediction Network (Student Model)

To predict the final ratings, we first concatenate the embeddings of the shared and private features as $\mathbf{s}_{ij} = [\bar{\mathbf{p}}_i, \bar{\mathbf{q}}_j]$ and $\mathbf{t}_{ij} = [\tilde{\mathbf{p}}_i, \tilde{\mathbf{q}}_j]$, respectively. Then the result is computed by

$$\hat{r}_{ij} = \text{PREDICT}(\mathbf{s}_{ij}, \mathbf{t}_{ij}), \quad (4)$$

$$\hat{r}_{ij} = \text{FM}([\mathbf{s}_{ij}, \mathbf{t}_{ij}]), \quad (5)$$

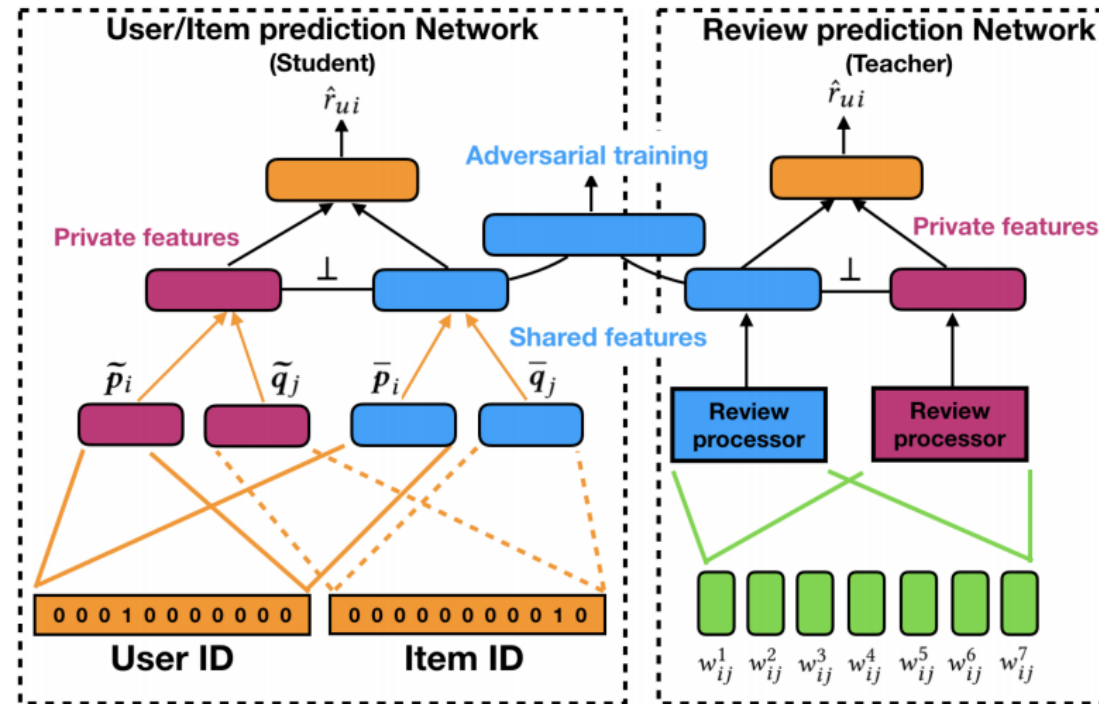
$$\text{FM}(\{z_i\}_{i=1}^n) = b_0 + \sum_{i=1}^n b_i z_i + \sum_{i=1}^n \sum_{j=(i+1)}^n (\mathbf{a}_i^T \mathbf{a}_j) z_i z_j, \quad (6)$$

Finally, the objective function to be minimized in a user-item prediction network is

$$L_{UI} = \frac{1}{|\mathcal{R}|} \sum_{(i,j) \in \mathcal{R}} \left((r_{ij} - \hat{r}_{ij})^2 + \lambda (\|\mathbf{s}_{ij}\|_2^2 + \|\mathbf{t}_{ij}\|_2^2) \right), \quad (7)$$

Review Prediction Network (Teacher Model)

we hope to transfer higher-quality review information into the student model, while the user/item irrelevant knowledge should be automatically filtered out.



$$\hat{r}_{w_{ij}} = V \cdot [s_{w_{ij}}, t_{w_{ij}}] + g,$$

$$L_{Review} = \frac{1}{|\mathcal{R}|} \sum_{(i,j) \in \mathcal{R}} (r_{ij} - \hat{r}_{w_{ij}})^2.$$

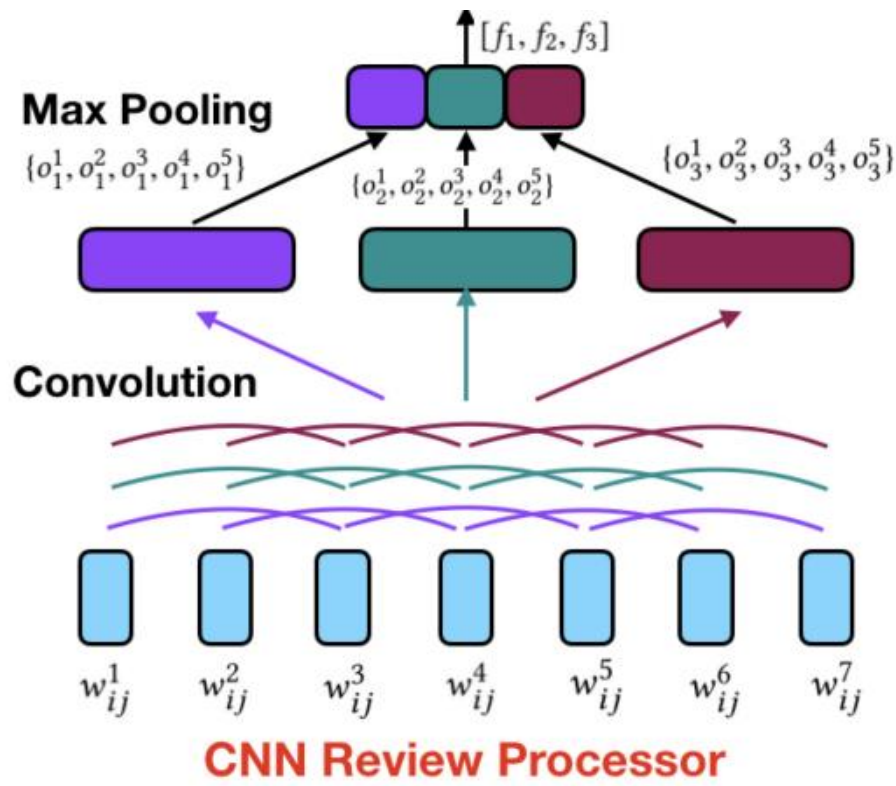


Fig. 5. Implementation of the CNN review processor.

$$o_j^l = \text{ReLU}(\mathbf{W}_{l:(l+t-1)} * \mathbf{K}_j + b_j),$$

$$f_j = \max\{o_j^1, o_j^2, \dots, o_j^{k-t+1}\},$$

$$\text{output} = h(\mathbf{H} \cdot [f_1, f_2, \dots, f_s]),$$

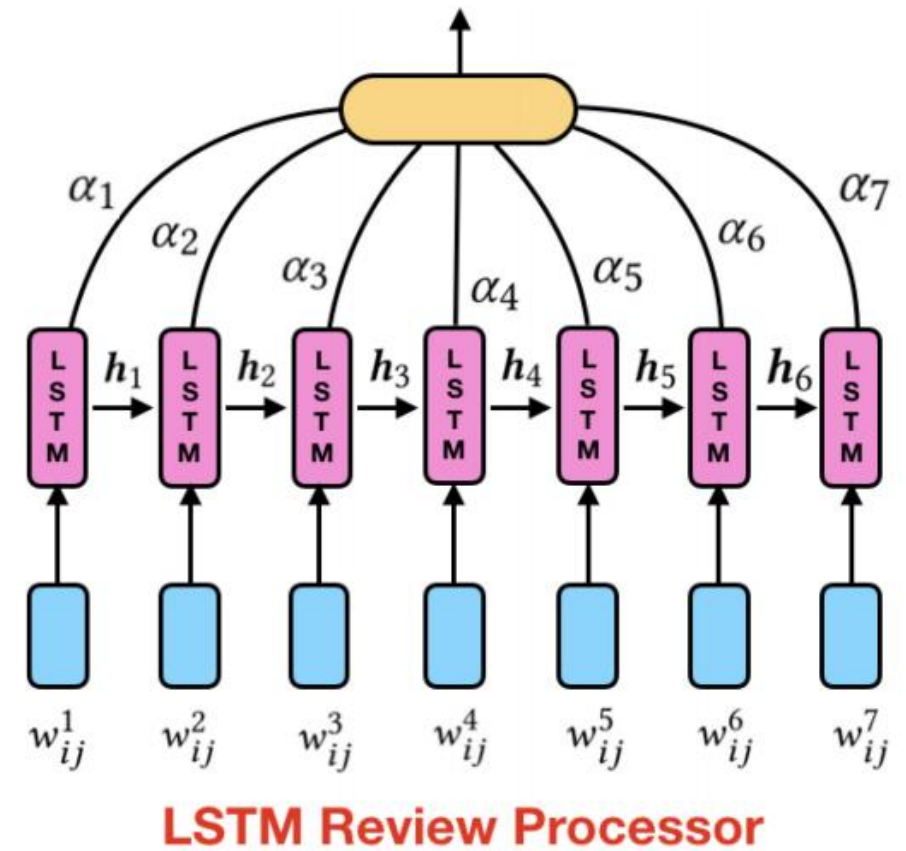


Fig. 6. Implementation of the LSTM review processor.

$$i_t = \sigma(V^i \cdot [h_{t-1}, W_t] + g^i),$$

$$f_t = \sigma(V^f \cdot [h_{t-1}, W_t] + g^f),$$

$$o_t = \sigma(V^o \cdot [h_{t-1}, W_t] + g^o),$$

$$\hat{c}_t = \tanh(V^c \cdot [h_{t-1}, W_t] + g^c),$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t,$$

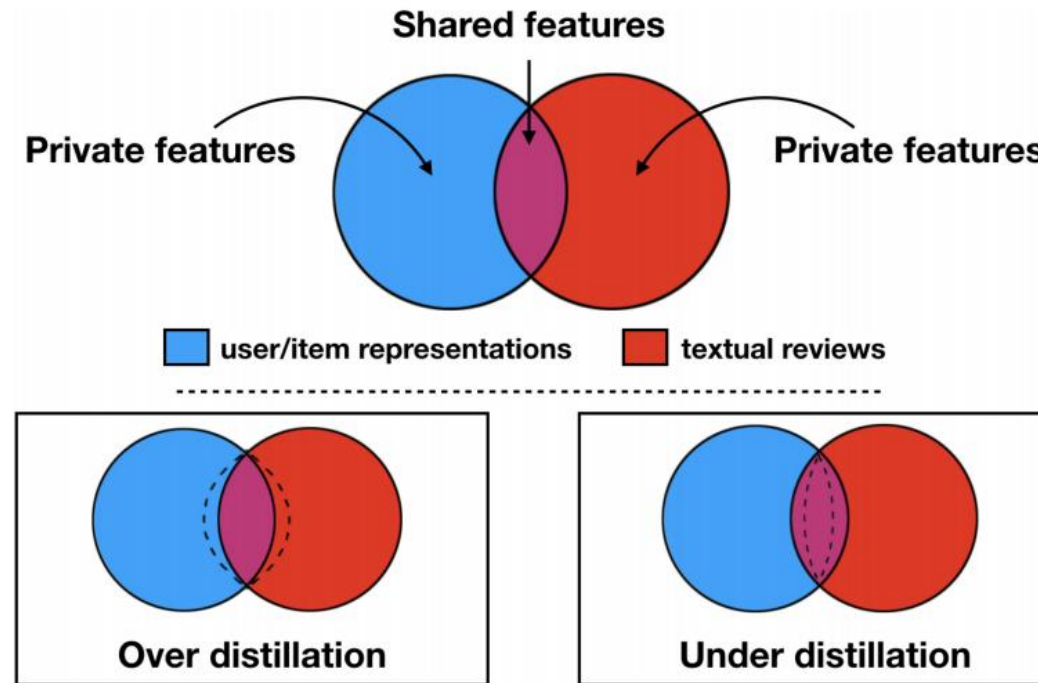
$$h_t = o_t \odot \tanh(c_t),$$

$$\alpha_t = \frac{\exp(\mathbf{e}^T \mathbf{h}_t)}{\sum_{j=1}^k \exp(\mathbf{e}^T \mathbf{h}_j)},$$

$$\text{output} = \sum_{t=1}^k \alpha_t \mathbf{h}_t.$$

Knowledge Distillation.

One is **over-distillation**, where some user/item irrelevant information is distilled through the shared embeddings, which will introduce noise to the learning of representations. Another is **under-distillation**, where the shared embeddings did not distill sufficient user/item relevant knowledge.



adversarial adaption between shared embeddings
orthogonality constraints between shared and private embeddings

adversarial adaption

In our model, the user-item prediction network (generator G) tries to generate s_{ij} that cannot be distinguished from s_w in the review prediction network by a discriminator D , which is trained to discover which network the shared embedding comes from. Formally, our model gives the following min-max objective function:

$$L_{GAN} = \min_{\Phi} \max_D \left(E_{w \sim W} [\log D(s_w)] + E_{(i,j) \sim S} [\log(1 - D(s_{ij}))] \right), \quad (22)$$

where W is the set of all reviews, Φ is the parameter set in user-item prediction network, and D is specialized as the sigmoid function.

solve the problem of over-distillation in Figure 2. **On one hand**, to fit the ground truth, the irrelevant information in the shared embedding will be penalized by the backpropagated supervision signal, which forces the shared embedding to only encode user/item relevant information. **On the other hand**, under the adversarial learning paradigm, the relevant information of the shared embedding is encouraged to be effectively distilled, such that the latent factors can learn more useful knowledge to predict user ratings.

Orthogonality Constraint

The above strategy tries to enforce that the shared embedding only encodes relevant information, however, such information may also exist in the private embeddings. To push the user-item related knowledge from the private embeddings into the shared ones, we introduce orthogonality constraint in our model.

Essentially, we hope the overlap between the shared and private embeddings to be as small as possible.

$$L_{Orth} = \frac{1}{|\mathcal{R}|} \sum_{(i,j) \in \mathcal{R}} \left(\mathbf{s}_{ij}^T \mathbf{t}_{ij} + \mathbf{s}_{\mathbf{w}_{ij}}^T \mathbf{t}_{\mathbf{w}_{ij}} \right),$$

Overall Objective Function

$$L = L_{UI} + \lambda_1 L_{Review} + \lambda_2 L_{GAN} + \lambda_3 L_{Tra} + \lambda_4 L_{Orth},$$

● Further Discussion

Comparison between LSTM and CNN Review Processor.

The difference is that the **CNN review processor** usually enables us to have higher training efficiency, while in **the LSTM review processor**, we can identify which words are more important according to their corresponding attention weights, which makes our model more transparent and explainable.

Comparison between SDNet and TransNet.

Comparison between SDNet and Generalized Distillation Framework

EXPERIMENTS

● Experimental Setup

Datasets

Table 3. Basic Statistics of Our Datasets

Datasets	#Users	#Items	#Words	#Total(R)	#Test(R)	#WPR	Density
Amazon Instant Video	5,130	1,685	85,480	37,126	3,729	93.55	0.433%
Automotive	2,928	1,835	46,400	20,473	1,828	86.96	0.381%
Baby	19,445	7,050	149,783	160,792	15,918	100.93	0.122%
Beauty	2,2363	12,101	175,974	198,502	22,569	90.58	0.071%
Cell Phones and Accessories	27,879	10,429	211,194	194,439	18,225	93.46	0.067%
Clothing Shoes and Jewelry	39,387	23,033	163,696	278,677	26,071	61.17	0.031%
Digital Music	5,541	3,568	143,566	64,706	6,109	204.73	0.327%
Grocery and Gourmet Food	1,4681	8,713	178,860	151,254	18,471	95.19	0.118%
Health and Personal Care	38,609	18,534	343,501	346,355	39,113	96.56	0.048%
Home and Kitchen	66,519	28,237	451,879	551,682	59,675	99.71	0.029%
Musical Instruments	1,429	900	30,988	10,261	918	92.28	0.798%
Office Products	4,905	2,420	126,230	53,258	6,575	148.46	0.449%
Patio Lawn and Garden	1,686	962	51,835	13,272	1,353	160.44	0.818%
Pet Supplies	19,856	8,510	153,947	157,836	16,294	89.95	0.093%
Sports and Outdoors	35,598	18,357	303,145	296,337	31,339	88.86	0.045%
Tools and Home Improvement	16,638	10,217	210,193	134,476	13,900	111.66	0.079%
Toys and Games	19,412	11,924	189,720	167,597	1,5481	101.90	0.072%
Video Games	24,303	10,672	556,049	231,780	29,585	210.55	0.089%
Raw Amazon Instant Video	426,910	23,962	334,179	583,914	7,173	54.34	0.0057%
Raw Digital Music	478,201	266,393	718,686	835,953	23,212	77.64	0.0007%
Raw Musical Instruments	338,967	83,025	539,873	499,730	8,504	89.88	0.0018%

Total(R) means the total number of records, Test(R) means the number of records used for testing, and WPR is the average number of words per review.

Evaluation Protocols

$$MSE = \frac{1}{|S|} \sum_{(i,j) \in S} (r_{ij} - \hat{r}_{ij})^2,$$

Baselines

- **MF**: This is a basic matrix factorization method [20]. The user-item rating matrix is estimated by the multiplication of two low-rank matrices.
- **PMF**: This method [34] generalizes MF into a probability form, where the user/item latent factors are assumed to follow Gaussian distribution. We learn the model parameters by stochastic gradient decent (SGD).
- **CTR**: This is a well-known probabilistic recommender model [43] leveraging textual information.
- **DeepCoNN**: This is the first deep model designed for review-based recommendation [51], where user reviews are modeled by the CNN.
- **TransNet, TransNet-Ext**: They are the state-of-the-art methods [3] for review-based recommendation. The difference between them is that TransNet only inputs all the reviews that belong to the current user/item, while TransNet-Ext further includes user/item latent factors in the source network. We implement them using the code provided by the authors.²

● Performance Comparison

Table 4. Comparison of MSE Between the Baselines and Our Models

Dataset	MF	PMF	CTR	DCN	Tran	TranE	SDNet		imp
							CNN	LSTM	
Amazon Instant Video	1.401	1.268	1.231	1.195	1.189	1.033*	0.988	0.994	4.3%
Automotive	1.303	1.026	1.020	0.986	0.977	0.960*	0.929	0.937	3.2%
Baby	1.371	1.231	1.229	1.220	1.203*	1.211	1.136	1.132	5.9%
Beauty	1.542	1.312	1.261	1.259	1.253	1.245*	1.214	1.219	2.4%
Cell Phones and Accessories	1.751	1.421	1.342	1.332	1.256*	1.283	1.226	1.230	2.4%
Clothing Shoes and Jewelry	1.821	1.312	1.291	1.250*	1.255	1.297	1.192	1.199	4.6%
Digital Music	1.511	1.290	1.211	1.170	1.113	1.036*	0.981	0.974	6.0%
Grocery and Gourmet Food	1.682	1.473	1.246	1.276	1.118	1.072*	1.037	1.044	3.2%
Health and Personal Care	1.571	1.259	1.244	1.207*	1.209	1.243	1.180	1.187	2.2%
Home and Kitchen	1.862	1.321	1.280	1.215	1.201	1.187*	1.141	1.144	3.9%
Musical Instruments	1.323	1.205	0.715	0.709	0.711	0.707*	0.681	0.678	4.1%
Office Products	0.991	0.901	0.891	0.887	0.876	0.785*	0.756	0.762	3.7%
Patio Lawn and Garden	1.491	1.256	1.183	1.147	1.161	1.004*	0.966	0.964	4.0%
Pet Supplies	1.422	1.305	1.311	1.347	1.304	1.271*	1.237	1.224	3.7%
Sports and Outdoors	0.996	0.971	0.961	0.961	0.943*	0.952	0.919	0.926	2.5%
Tools and Home Improvement	1.431	1.223	1.112	1.102	1.056	0.988*	0.960	0.952	3.6%
Toys and Games	1.256	1.149	0.981	0.910	1.024	0.901*	0.878	0.876	2.7%
Video Games	1.543	1.512	1.358	1.268*	1.286	1.273	1.199	1.190	6.2%
Raw Amazon Instant Video	1.421	1.341	1.201	1.162	1.143	1.139*	1.091	1.110	4.2%
Raw Digital Music	1.001	0.931	0.881	0.821	0.811	0.801*	0.785	0.792	2.0%
Raw Musical Instruments	1.321	1.223	1.104	1.079	1.074*	1.083	1.046	1.053	2.6%

- the importance of textual features for rating prediction in the field of recommender systems
- our model (the better version between CNN and LSTM review processors) can consistently obtain the best performance on both dense and sparse datasets, it can produce more useful and clean information to help the user/item latent factors to fit the ground-truth, which finally leads to better performance
- The reason can be that long reviews may contain more useless or even noisy text, and with the help of the selective distillation mechanism, our model can filter this noisy information more effectively, which leads to better performance
- Interestingly, we find that the difference between the CNN and LSTM review processors was not obvious on most datasets. The reason may be that although LSTM is good at capturing word sequential information

● Parameter Analysis

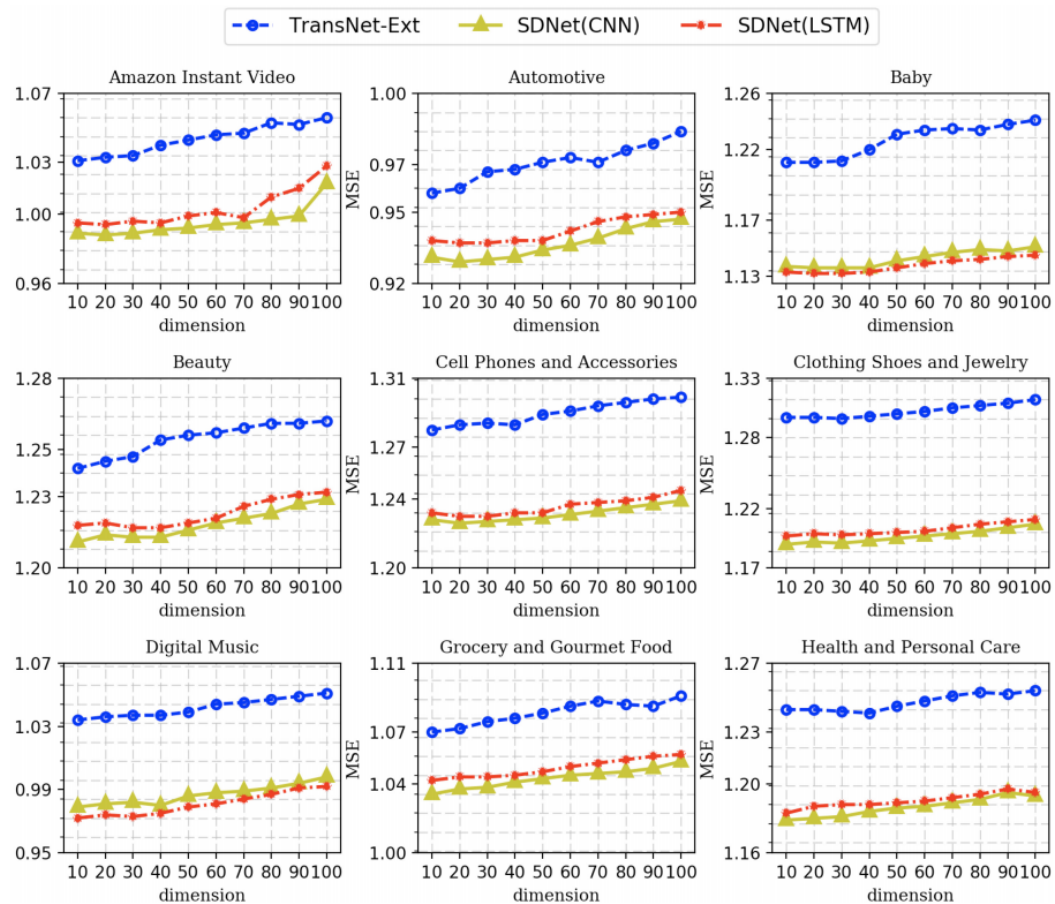


Fig. 7. Influence of the latent factor dimension.

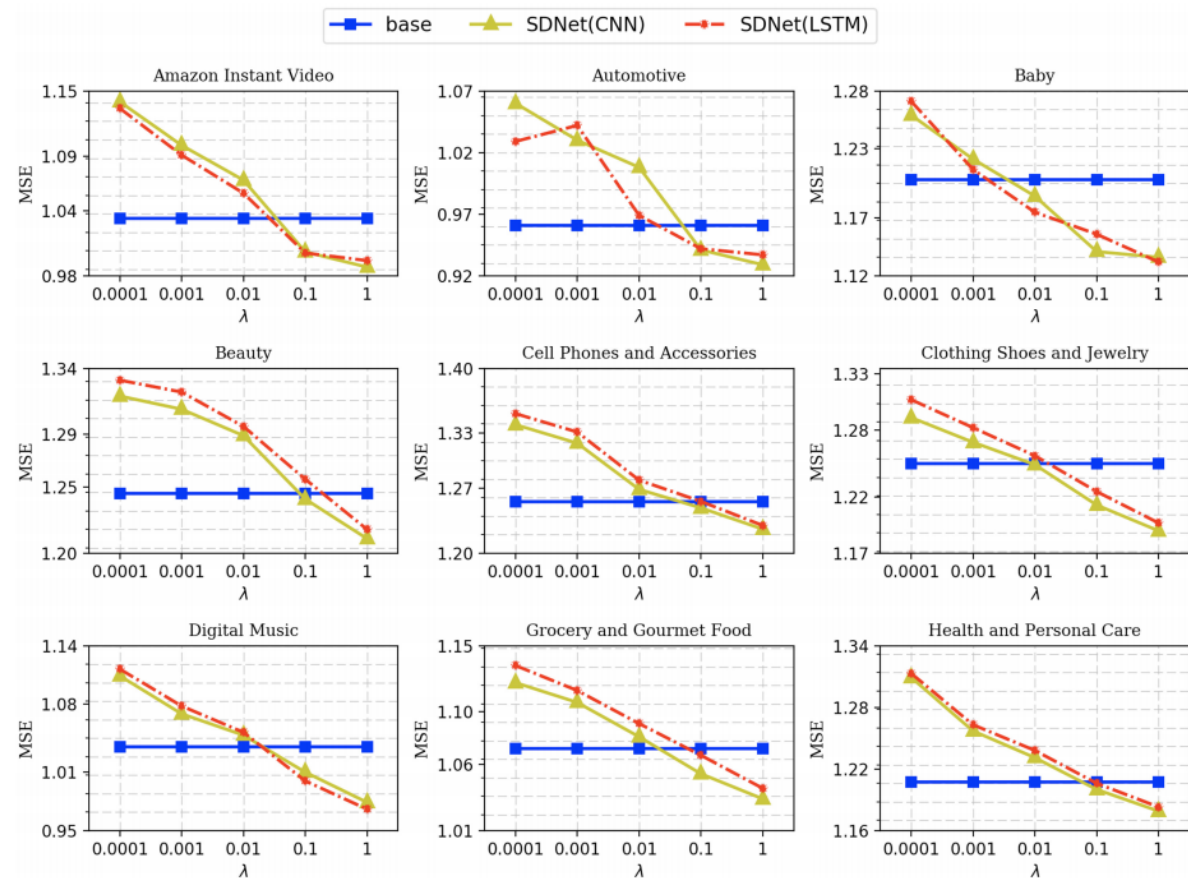


Fig. 8. Influence of the regularizer coefficient.

● Efficiency Comparison

Table 5. Efficiency Comparison

Dataset	PMF	DeepCoNN	TransNet	TransNet-Ext	SDNet
Amazon Instant Video	1.825s	17.904s	18.480s	17.762s	1.939s
Automotive	0.911s	3.609s	3.671s	3.521s	0.953s
Cell Phones and Accessories	3.004s	78.539s	79.237s	84.530s	3.126s
Digital Music	5.212s	89.712s	89.506s	91.119s	5.694s
Grocery and Gourmet Food	8.945s	93.462s	94.450s	93.218s	9.167s
Musical Instruments	0.483s	1.935s	1.943s	1.955s	0.544s
Office Products	1.452s	30.176s	35.229s	34.975s	1.521s
Patio Lawn and Garden	0.302s	5.427s	5.671s	5.891s	0.360s

We do not distinguish CNN and LSTM versions of SDNet because they share the same model at test time.

● Model Ablation: Effect of Different Components in SDNet

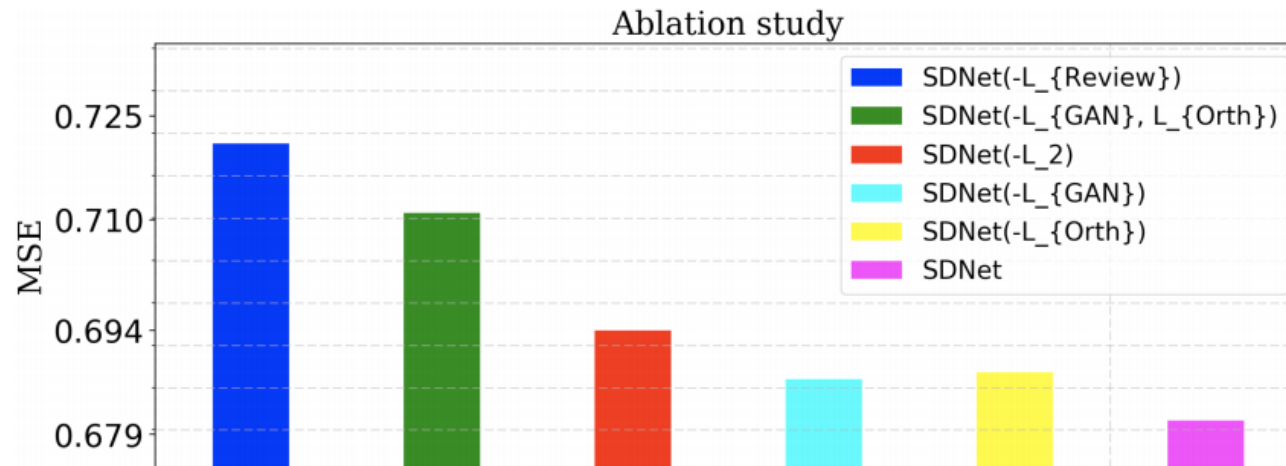
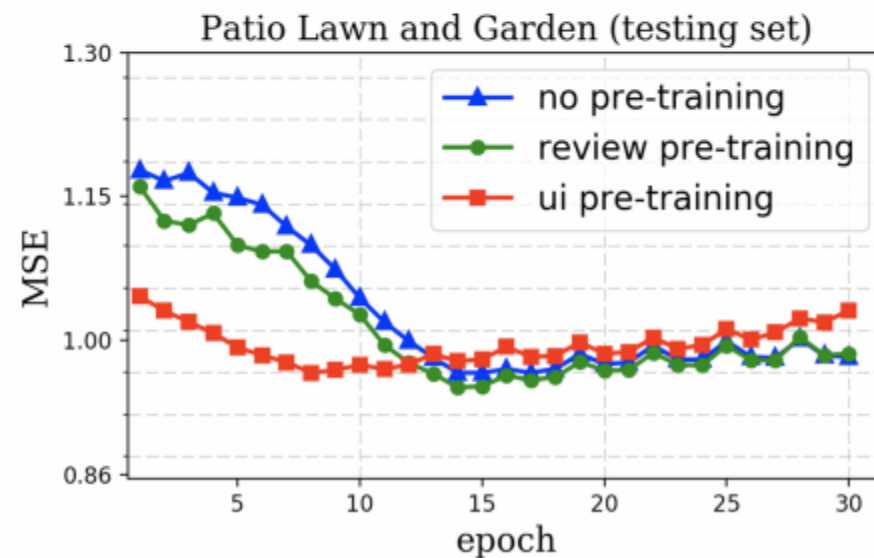
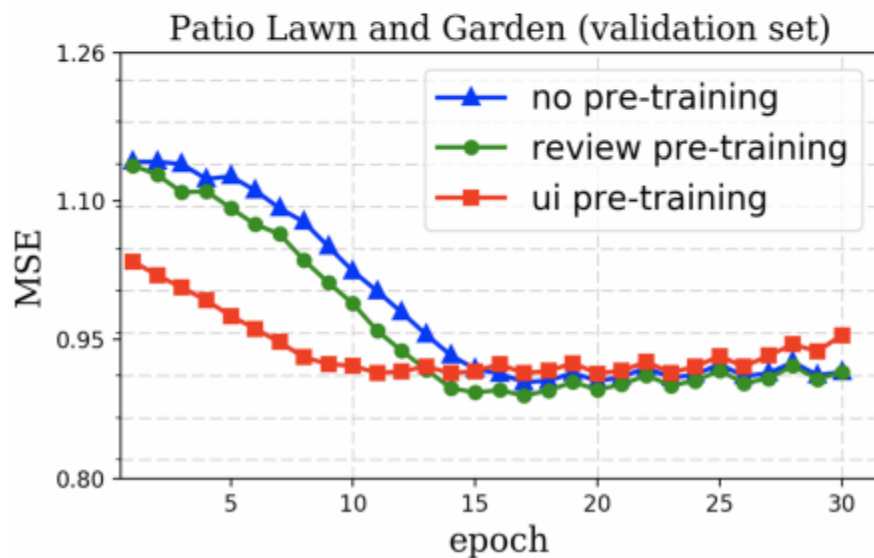
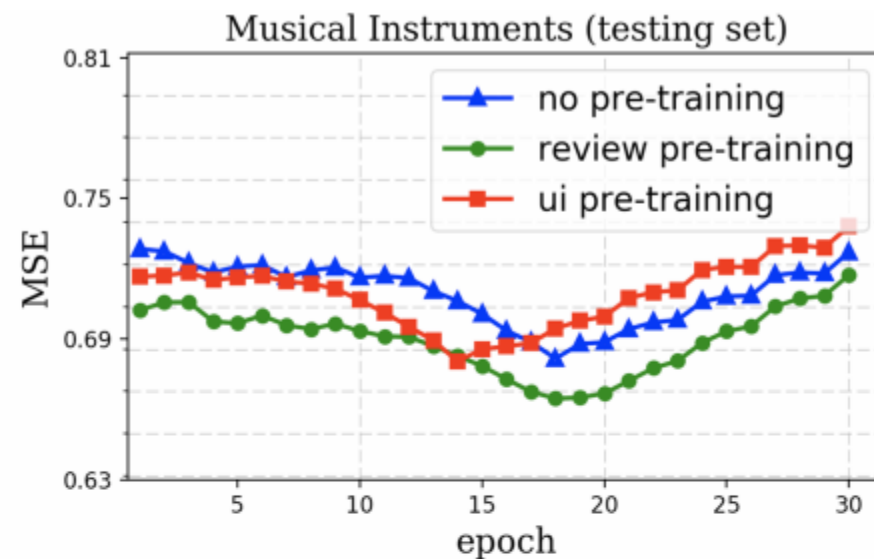
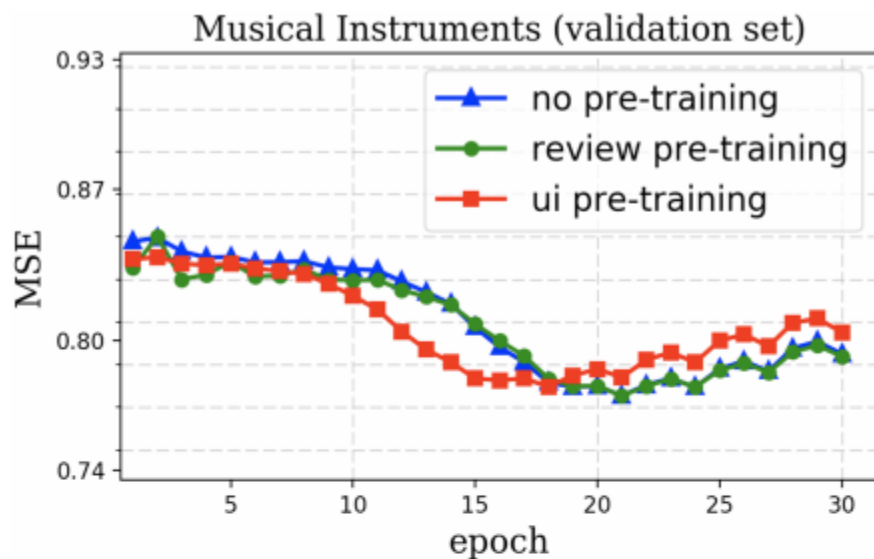


Fig. 9. The effect of different components in our model.

● Model Pre-Training



● Visualization of the Distilled Knowledge

Table 6. Visualization of the Distilled Knowledge

Rating	Review
5.0	I normally get elixir strings but these sounded too good to <i>pass</i> up. They are <u>indeed</u> some <u>high</u> tech <i>strings</i> with <i>features</i> including <i>break resistance</i> and improved tuning stability. They feel and sound <u>good</u> so grab a pack and get jamming you'll <u>definitely want</u> to check these out.
5.0	I have a pretty crappy <u>ear</u> so i rely on it to <i>tune</i> <u>correctly</u> responds <u>fast</u> display is incredible. I wish my car dash looked <u>like</u> this thing <u>fits</u> easily on the headstock of the guitar in front or behind tunes by <i>sensing</i> the <i>vibration</i> of the instrument not <i>using</i> a <i>microphone</i> .
2.0	To star after <i>returning</i> it back to <u>brookmays</u> . The seller I am <u>really</u> <u>disappointed</u> by the seller brookmays' <u>ignorant</u> <i>responses</i> to my return of the <u>defective</u> guitar pickup. After two weeks of the returned item was delivered i didn't <i>hear anything</i> from the seller <i>regarding</i> the return.
4.0	This amp cable is <u>perfect</u> . It fits my need for practicing. It's just the right length for sitting on the <i>couch strumming daily</i> . The cable also <i>looks</i> <u>nice</u> and the fender look I <i>would</i> <u>highly recommend</u> this and give it a no-brainer 5 star <u>rating</u> . I have nothing negative to say about this.
3.0	Looking for options around the old dropped <u>pick syndrome</u> . I was <u>skeptical</u> as the <u>pick</u> at first glance seems <u>slicker</u> than <i>most</i> , however, once <i>between thumb</i> and <i>index finger</i> these hang on.

We present five pieces of user review in the musical instruments dataset. The words most related to the shared and private embeddings are labeled by underlined and italic fonts, respectively.

CONCLUSION

- we proposed to formulate the problem of recommendation with external knowledge into a generalized distillation framework.
- We take user reviews as the external knowledge, and further developed a SDNet to transfer informative review signals from the teacher model into the student model for effective user/item representation learning.
- We designed two key strategies for knowledge distillation, adversarial adaption and orthogonality constraint, to guarantee the quality of the knowledge distilled between different networks.
- Extensive experiments verified that our model can significantly outperform many state-of-the-art methods in terms of both effectiveness and efficiency perspectives.

谢 谢