An Attentive Inductive Bias for Sequential Recommendation beyond the Self-Attention

Yehjin Shin*, Jeongwhan Choi*, Hyowon Wi, Noseong Park

Yonsei University, Seoul, South Korea
{yehjin.shin, jeongwhan.choi, wihyowon, noseong}@yonsei.ac.kr

code: https://github.com/yehjin-shin/BSARec.

AAAI 2024

Reported by Minqin Li

Introduction

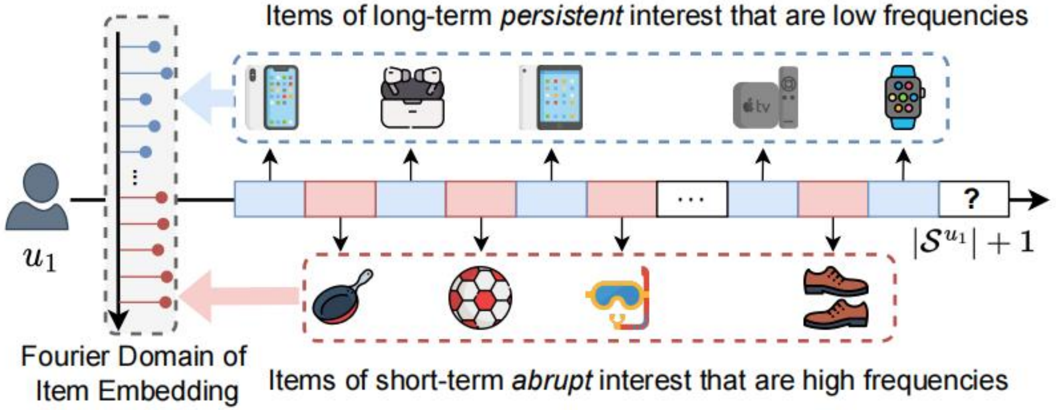
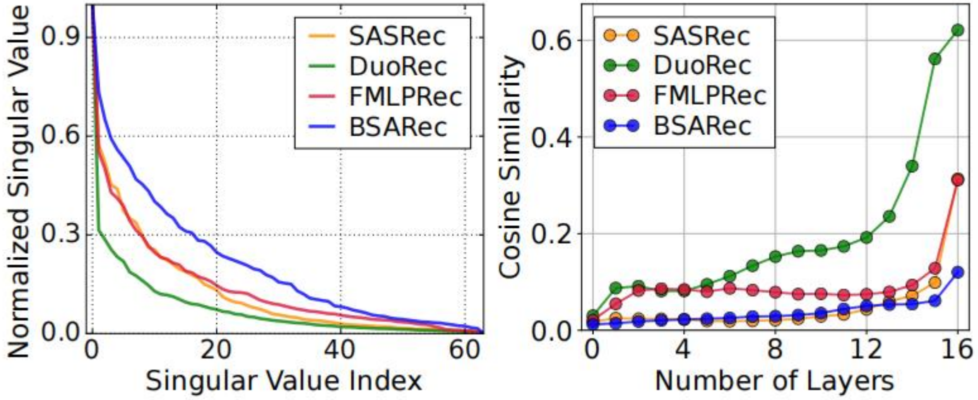
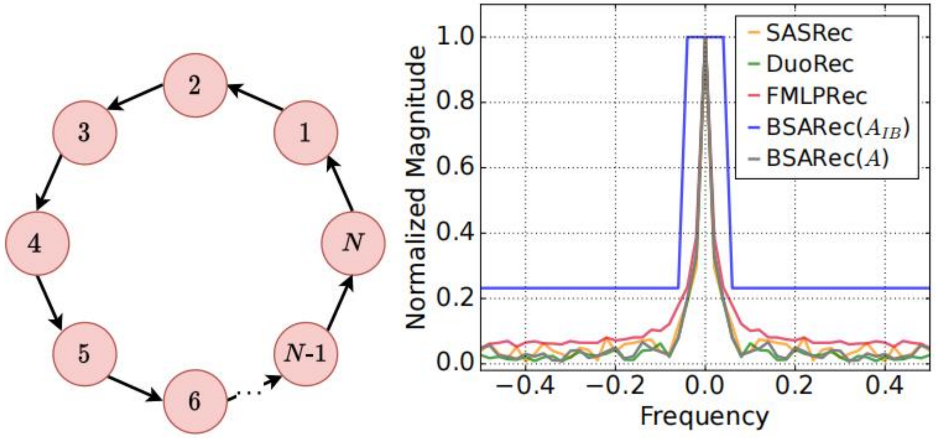


Figure 1: Illustration of high and low-frequency signals in SR.



(a) Singular value (b) Cosine similarity

Figure 3: Visualization of oversmoothing in LastFM. The singular values and cosine similarity of user sequence output embedding.



(a) A ring graph (b) Spectral responses

Figure 2: (a) A ring graph with N nodes, and (b) visualization of the filter of the self-attentions in LastFM.

The self-attention causes the oversmoothing problem that Tranformer-based SR models lose feature representation in deep layers.This inevitably causes the model to fail to capture the user's detailed preferences, and performance degradation is a natural result.

We not only alleviate oversmoothing using a high-pass filter as motivation against this background, but also try to capture short-term preferences of user behavior patterns through inductive bias.

Method

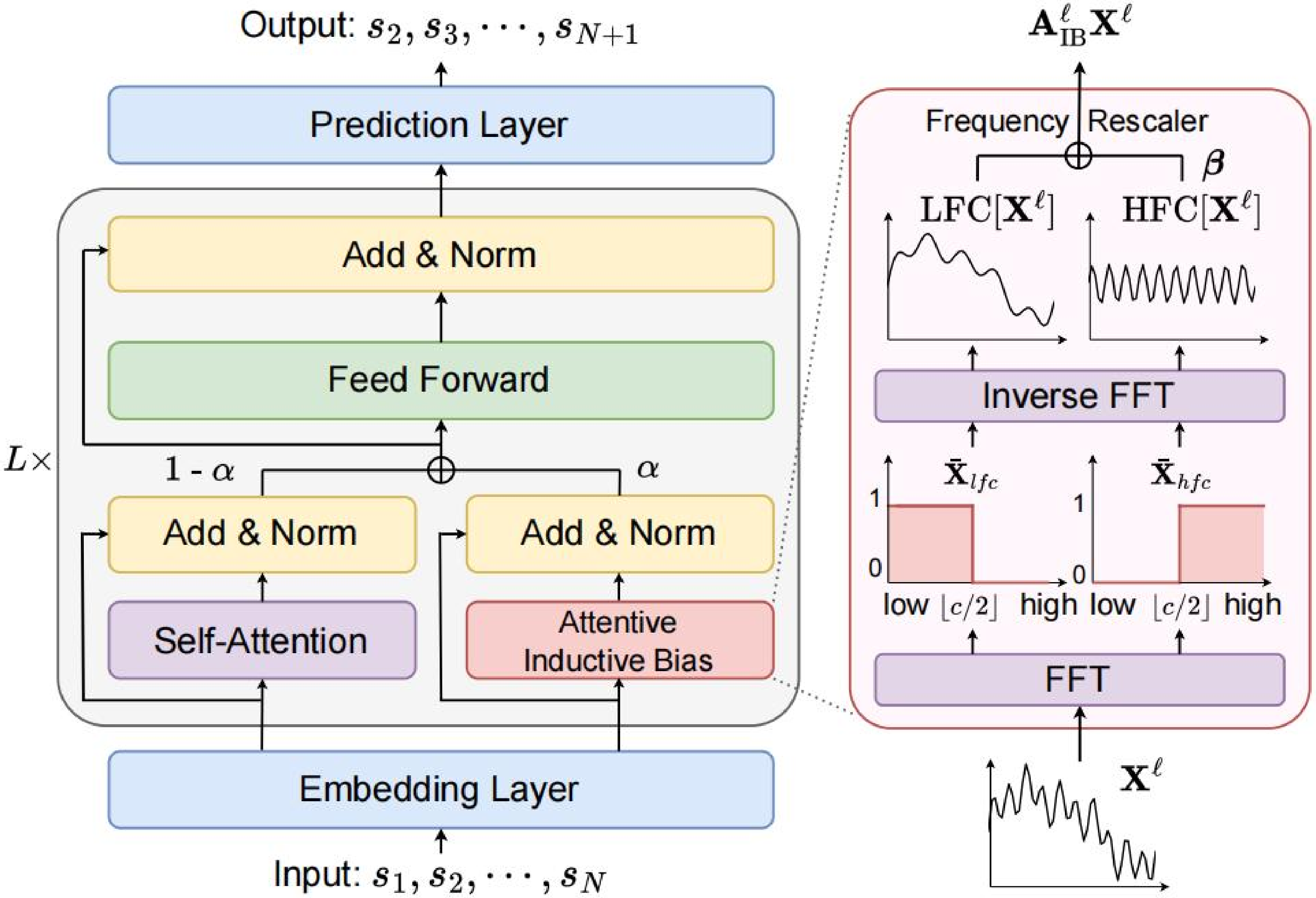
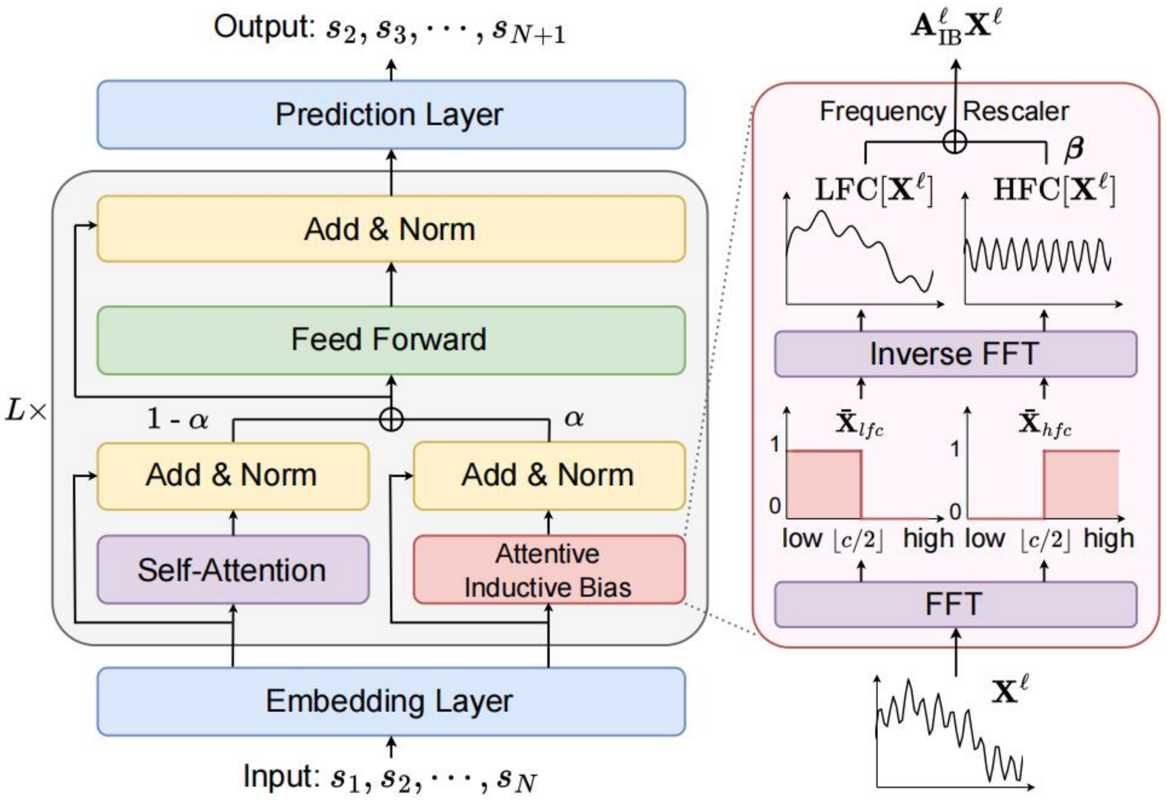


Figure 4: Architecture of our proposed BSARec.

Method



$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{QK}^{\mathrm{T}}}{\sqrt{d}}\right) \tag{1}$$

$$\mathrm{LFC}[\boldsymbol{x}] = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_c]\,\bar{\boldsymbol{x}}_{\mathrm{lfc}} \in \mathbb{R}^N \tag{2}$$

$$\mathrm{HFC}[\boldsymbol{x}] = [\boldsymbol{f}_{c+1}, \boldsymbol{f}_{c+2}, \ldots, \boldsymbol{f}_N]\,\bar{\boldsymbol{x}}_{\mathrm{hfc}} \in \mathbb{R}^N \tag{3}$$

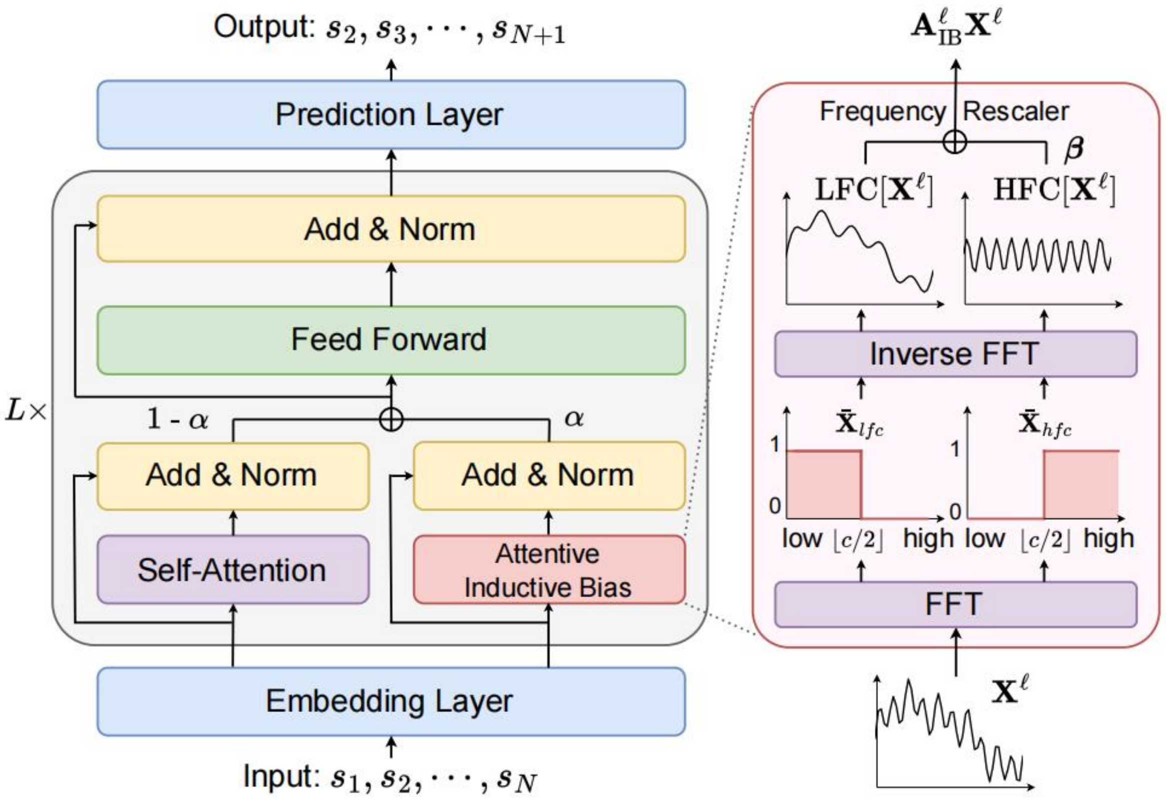
$$\mathbf{E}^u = \mathrm{Dropout}(\mathrm{LayerNorm}(\mathbf{E}^u + \mathbf{P})) \tag{4}$$

$$\mathbf{S}^{\ell} = \widetilde{\mathbf{A}}^{\ell}\mathbf{X}^{\ell} = \alpha\mathbf{A}_{\mathrm{IB}}^{\ell}\mathbf{X}^{\ell} + (1-\alpha)\mathbf{A}^{\ell}\mathbf{X}^{\ell} \tag{5}$$

$$\widehat{\mathbf{X}}^{\ell} = \mathrm{MSA}(\mathbf{X}^{\ell}) = [\mathbf{S}^1, \mathbf{S}^2, \ldots \mathbf{S}^h]\mathbf{W}^O \tag{6}$$

$$\mathbf{A}_{\mathrm{IB}}^{\ell}\mathbf{X}^{\ell} = \mathrm{LFC}[\mathbf{X}^{\ell}] + \beta\mathrm{HFC}[\mathbf{X}^{\ell}] \tag{7}$$

Method



$$\widetilde{\mathbf{X}}^{\ell} = (\mathrm{GELU}(\widehat{\mathbf{X}}^{\ell}\mathbf{W}_1^{\ell} + \mathbf{b}_1^{\ell}))\mathbf{W}_2^{\ell} + \mathbf{b}_2^{\ell} \quad (8)$$

$$\mathbf{X}^{\ell+1} = \mathrm{LayerNorm}(\mathbf{X}^{\ell} + \widehat{\mathbf{X}}^{\ell} + \mathrm{Dropout}(\widetilde{\mathbf{X}}^{\ell})) \quad (9)$$

$$\hat{y}_i = p(v^u_{|\mathcal{S}^u|+1} = v|\mathcal{S}^u) = \mathbf{e}_v^{\mathsf{T}}\mathbf{X}^L_{|\mathcal{S}^u|} \quad (10)$$

$$\mathcal{L} = -log\frac{\exp(\hat{y}_g)}{\sum_{i\in|\mathcal{V}|}\exp(\hat{y}_i)} \quad (11)$$

Experiments

Method	Inductive Bias	Self-Attention	High-pass Filter
SASRec	✗	✓	✗
BERT4Rec	✗	✓	✗
FMLPRec	✓	✗	✗
DuoRec	✗	✓	✗
BSARec	✓	✓	✓

Table 1: Comparison of existing Transformer-based methods that differ at three points: i) using inductive bias, ii) using self-attentions, and iii) using high-pass filters

Experiments

	# Users	# Items	# Interactions	Avg. Length	Sparsity
Beauty	22,363	12,101	198,502	8.9	99.93%
Sports	25,598	18,357	296,337	8.3	99.95%
Toys	19,412	11,924	167,597	8.6	99.93%
Yelp	30,431	20,033	316,354	10.4	99.95%
LastFM	1,090	3,646	52,551	48.2	98.68%
ML-1M	6,041	3,417	999,611	165.5	95.16%

Table 5: Statistics of the processed datasets

Experiments

Datasets	Metric	Caser	GRU4Rec	SASRec	BERT4Rec	FMLPRec	DuoRec	FEARec	BSARec	Improv.
Beauty	HR@5	0.0125	0.0169	0.0340	0.0469	0.0346	0.0707	0.0706	**0.0736**	4.10%
	HR@10	0.0225	0.0304	0.0531	0.0705	0.0559	0.0965	0.0982	**0.1008**	2.65%
	HR@20	0.0403	0.0527	0.0823	0.1073	0.0869	0.1313	0.1352	**0.1373**	1.55%
	NDCG@5	0.0076	0.0104	0.0221	0.0311	0.0222	0.0501	0.0512	**0.0523**	2.15%
	NDCG@10	0.0108	0.0147	0.0283	0.0387	0.0291	0.0584	0.0601	**0.0611**	1.66%
	NDCG@20	0.0153	0.0203	0.0356	0.0480	0.0369	0.0671	0.0694	**0.0703**	1.30%
Sports	HR@5	0.0091	0.0118	0.0188	0.0275	0.0220	0.0396	0.0411	**0.0426**	3.65%
	HR@10	0.0163	0.0187	0.0298	0.0428	0.0336	0.0569	0.0589	**0.0612**	3.90%
	HR@20	0.0260	0.0303	0.0459	0.0649	0.0525	0.0791	0.0836	**0.0858**	2.63%
	NDCG@5	0.0056	0.0079	0.0124	0.0180	0.0146	0.0276	0.0286	**0.0300**	4.90%
	NDCG@10	0.0080	0.0101	0.0159	0.0229	0.0183	0.0331	0.0343	**0.0360**	4.96%
	NDCG@20	0.0104	0.0131	0.0200	0.0284	0.0231	0.0387	0.0405	**0.0422**	4.20%
Toys	HR@5	0.0095	0.0121	0.0440	0.0412	0.0432	0.0770	0.0783	**0.0805**	2.81%
	HR@10	0.0161	0.0211	0.0652	0.0635	0.0671	0.1034	0.1054	**0.1081**	2.56%
	HR@20	0.0268	0.0348	0.0929	0.0939	0.0974	0.1369	0.1397	**0.1435**	2.72%
	NDCG@5	0.0058	0.0077	0.0297	0.0282	0.0288	0.0568	0.0574	**0.0589**	2.61%
	NDCG@10	0.0079	0.0106	0.0366	0.0353	0.0365	0.0653	0.0661	**0.0679**	2.72%
	NDCG@20	0.0106	0.0140	0.0435	0.0430	0.0441	0.0737	0.0747	**0.0768**	2.81%
Yelp	HR@5	0.0117	0.0130	0.0149	0.0256	0.0159	0.0271	0.0262	**0.0275**	1.48%
	HR@10	0.0197	0.0221	0.0249	0.0433	0.0287	0.0442	0.0437	**0.0465**	5.20%
	HR@20	0.0337	0.0383	0.0424	0.0717	0.0490	0.0717	0.0691	**0.0746**	4.04%
	NDCG@5	0.0070	0.0080	0.0091	0.0159	0.0100	**0.0170**	0.0165	**0.0170**	0.00%
	NDCG@10	0.0096	0.0109	0.0123	0.0216	0.0142	0.0225	0.0221	**0.0231**	2.67%
	NDCG@20	0.0131	0.0150	0.0167	0.0287	0.0192	0.0294	0.0285	**0.0302**	2.72%
LastFM	HR@5	0.0303	0.0312	0.0413	0.0294	0.0367	0.0431	0.0431	**0.0523**	21.35%
	HR@10	0.0431	0.0404	0.0633	0.0459	0.0560	0.0624	0.0587	**0.0807**	27.49%
	HR@20	0.0642	0.0541	0.0927	0.0596	0.0826	0.0963	0.0826	**0.1174**	21.91%
	NDCG@5	0.0227	0.0217	0.0284	0.0198	0.0243	0.0300	0.0304	**0.0344**	13.16%
	NDCG@10	0.0268	0.0245	0.0355	0.0252	0.0306	0.0361	0.0354	**0.0435**	20.50%
	NDCG@20	0.0321	0.0280	0.0429	0.0286	0.0372	0.0446	0.0414	**0.0526**	17.94%
ML-1M	HR@5	0.0927	0.1005	0.1374	0.1512	0.1316	0.1838	0.1834	**0.1944**	5.77%
	HR@10	0.1556	0.1657	0.2137	0.2346	0.2065	0.2704	0.2705	**0.2757**	1.92%
	HR@20	0.2488	0.2664	0.3245	0.3440	0.3137	0.3738	0.3714	**0.3884**	3.91%
	NDCG@5	0.0592	0.0619	0.0873	0.1021	0.0846	0.1252	0.1236	**0.1306**	4.31%
	NDCG@10	0.0795	0.0828	0.1116	0.1289	0.1087	0.1530	0.1516	**0.1568**	2.48%
	NDCG@20	0.1028	0.1081	0.1395	0.1564	0.1356	0.1790	0.1771	**0.1851**	3.41%

Experiments

Methods	Beauty		Toys	
	HR@20	NDCG@20	HR@20	NDCG@20
BSARec	**0.1373**	**0.0703**	**0.1435**	**0.0768**
Only $\mathbf{A}$	0.1265	0.0657	0.1320	0.0720
Only $\mathbf{A}_{\text{IB}}$	0.1338	0.0677	0.1402	0.0744
Scalar β	0.1333	0.0685	**0.1435**	0.0756

Table 3: Ablation studies on $\tilde{\mathbf{A}}$ and β. More results in other datasets are in Appendix.

Methods	Beauty		Sports		Toys		Yelp		LastFM		ML-1M	
	HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20
BSARec	**0.1373**	**0.0703**	**0.0858**	**0.0422**	**0.1435**	**0.0768**	**0.0746**	**0.0302**	**0.1174**	**0.0526**	**0.3884**	**0.1851**
Only $\mathbf{A}$	0.1265	0.0657	0.0779	0.0382	0.1320	0.0720	0.0618	0.0248	0.0899	0.0430	0.3826	0.1846
Only $\mathbf{A}_{\text{IB}}$	0.1338	0.0677	0.0857	0.0416	0.1402	0.0744	0.0705	0.0287	0.1009	0.0455	0.3780	0.1807
Scalar β	0.1333	0.0685	0.0838	0.0405	**0.1435**	0.0756	0.0707	0.0291	0.1092	0.0497	0.3762	0.1794

Table 7: Ablation on all datasets

Experiments

Methods	Beauty		ML-1M	
	# params	s/epoch	# params	s/epoch
BSARec	878,208	12.75	322,368	20.73
SASRec	877,824	10.41	321,984	19.37
DuoRec	877,824	19.26	321,984	32.33
FEARec	877,824	156.83	321,984	278.24

Table 4: The number of parameters and training time (runtime per epoch) on Beauty and ML-1M. More results in other datasets are in Appendix.

Methods	Beauty		Sports		Toys		Yelp		LastFM		ML-1M	
	# params	s/epoch	# params	s/epoch	# params	s/epoch	# params	s/epoch	# params	s/epoch	# params	s/epoch
BSARec	878,208	12.75	1,278,592	18.58	866,880	11.63	1,385,856	21.20	337,088	3.11	322,368	20.73
SASRec	877,824	10.41	1,278,208	15.32	866,496	9.96	1,385,472	18.25	336,704	2.80	321,984	19.37
DuoRec	877,824	19.26	1,278,208	27.99	866,496	18.79	1,385,472	31.08	336,704	4.24	321,984	32.33
FEARec	877,824	156.83	1,278,208	233.42	866,496	132.43	1,385,472	257.56	336,704	27.82	321,984	278.24

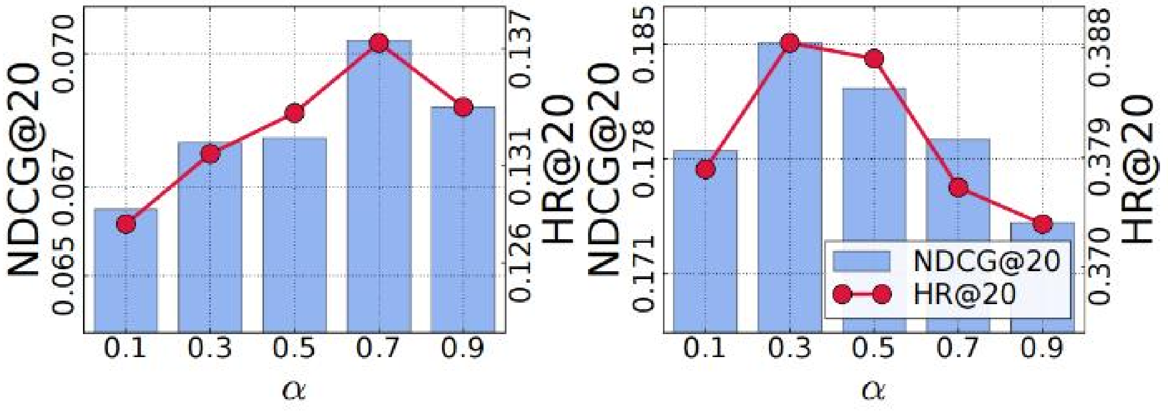
Table 8: The number of parameters and training time (runtime per epoch) on all datasets

Experiments

	Beauty	Sports	Toys	Yelp	LastFM	ML-1M
α	0.7	0.3	0.7	0.7	0.9	0.3
c	5	5	3	3	3	9
h	1	4	1	4	1	4
lr	5×10^{-4}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	5×10^{-4}

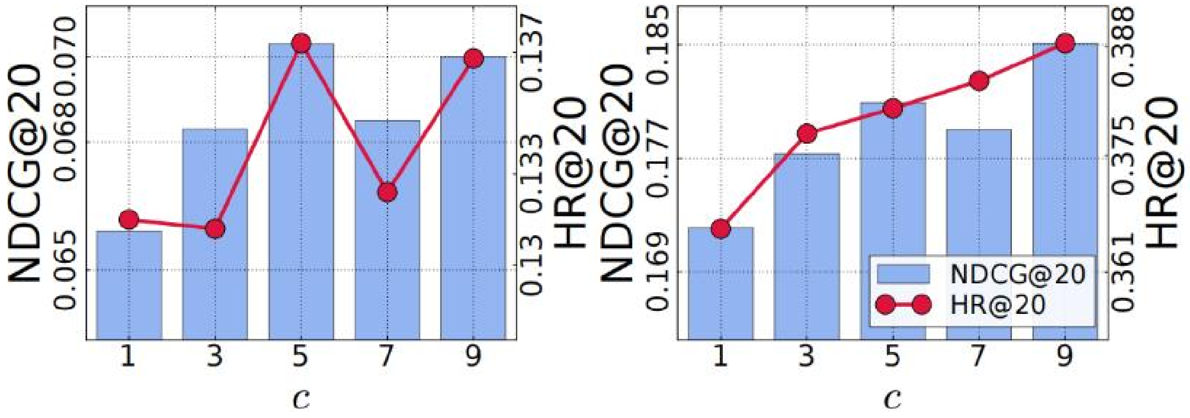
Table 6: Best hyperparameters of BSARec on all datasets

Experiments



(a) Beauty (b) ML-1M

Figure 5: Sensitivity to α. More results in other datasets are in Appendix.



(a) Beauty (b) ML-1M

Figure 6: Sensitivity to c. More results in other datasets are in Appendix.

Experiments

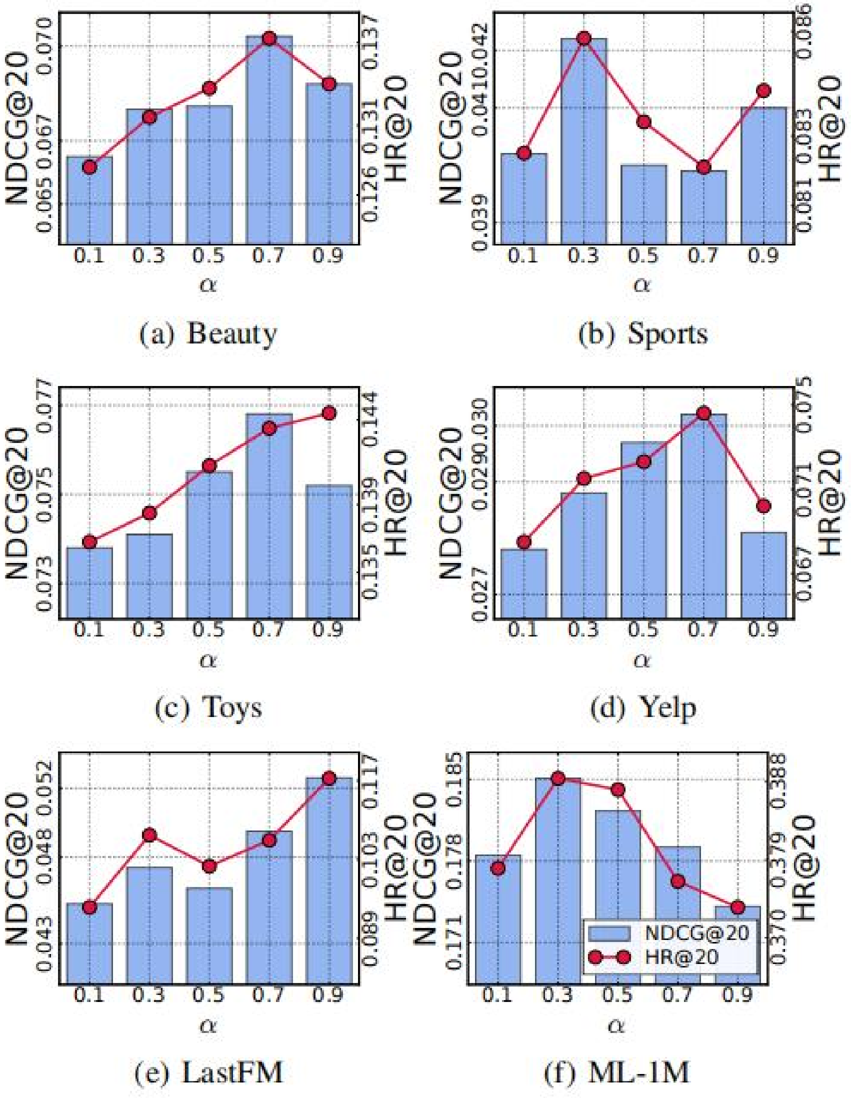


Figure 8: Sensitivity to α on all datasets

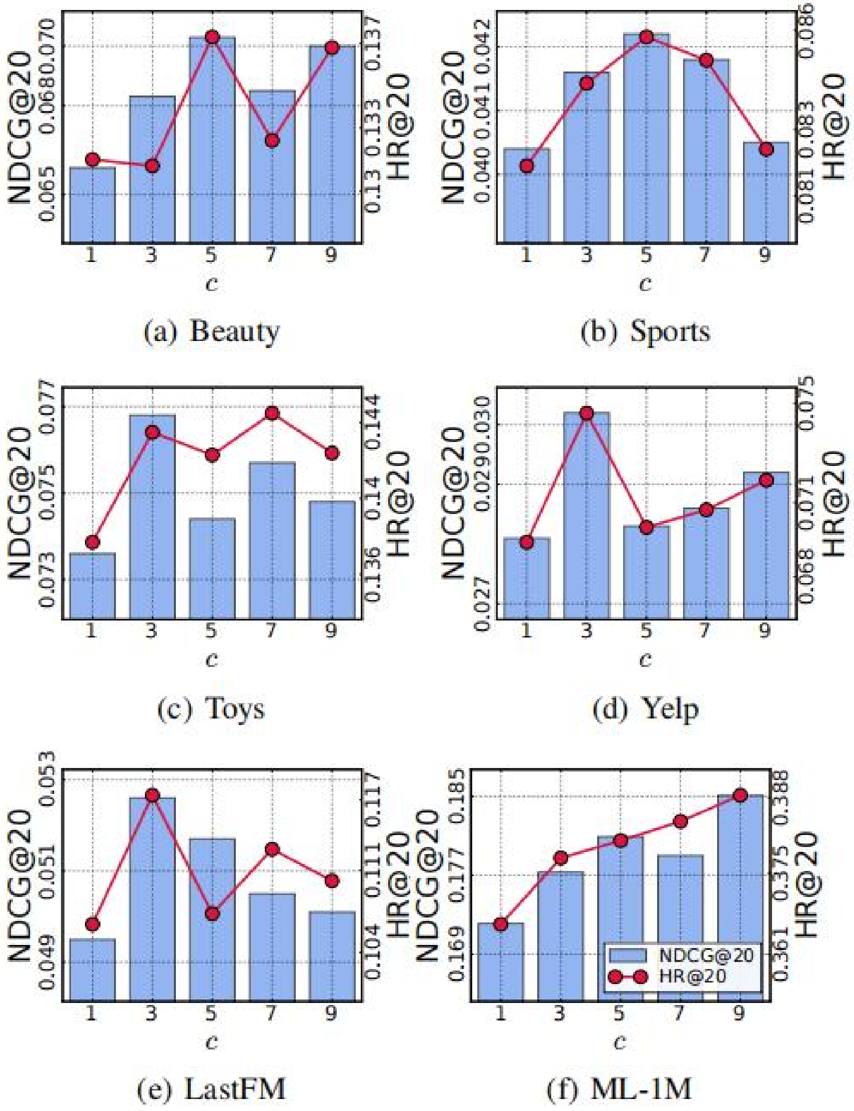
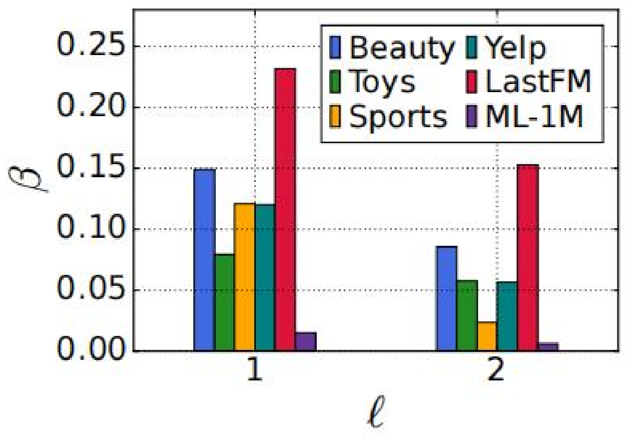
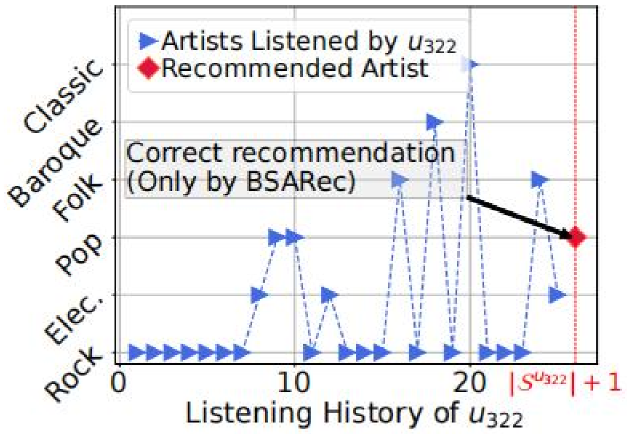


Figure 9: Sensitivity to c on all datasets



(a) Visualization of learned β



(b) Case study

Figure 7: (a) Visualization of learned β, and (b) an example recommendation in LastFM. The y-axis represents the genre of the artist the user listened to.

Thanks

Experiments

Dataset	Metric	SASRec	FMLPRec	BSARec
Beauty	HR@5	0.3512	0.3922	**0.4312**
	HR@10	0.4434	0.4914	**0.5225**
	NDCG@5	0.2628	0.2964	**0.3379**
	NDCG@10	0.2926	0.3284	**0.3673**
	MRR	0.2637	0.2949	**0.3350**
Sports	HR@5	0.3480	0.3781	**0.4133**
	HR@10	0.4717	0.4997	**0.5303**
	NDCG@5	0.2492	0.2739	**0.3102**
	NDCG@10	0.2891	0.3131	**0.3479**
	MRR	0.2520	0.2742	**0.3089**
Toys	HR@5	0.3594	0.3867	**0.4224**
	HR@10	0.4566	0.4852	**0.5180**
	NDCG@5	0.2726	0.2926	**0.3351**
	NDCG@10	0.3040	0.3244	**0.3659**
	MRR	0.2746	0.2917	**0.3349**
Yelp	HR@5	0.5553	0.6058	**0.6447**
	HR@10	0.7406	0.7707	**0.7848**
	NDCG@5	0.3902	0.4337	**0.4824**
	NDCG@10	0.4504	0.4873	**0.5280**
	MRR	0.3748	0.4114	**0.4587**
LastFM	HR@5	0.2716	0.2853	**0.3752**
	HR@10	0.3972	0.4138	**0.5028**
	NDCG@5	0.1871	0.1975	**0.2634**
	NDCG@10	0.2276	0.2394	**0.3045**
	MRR	0.1976	0.2081	**0.2636**
ML-1M	HR@5	0.6874	0.6763	**0.7023**
	HR@10	0.7904	0.7858	**0.7978**
	NDCG@5	0.5308	0.5212	**0.5646**
	NDCG@10	0.5642	0.5568	**0.5955**
	MRR	0.5020	0.4941	**0.5406**

Table 9: Performance comparison on 99 negative sampling

Experiments

$$\mathbf{F} = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & e^{2\pi i} & \dots & e^{2\pi i(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{2\pi i(j-1)\cdot 1} & \dots & e^{2\pi i(j-1)\cdot(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{2\pi i(N-1)} & \dots & e^{2\pi i(N-1)^2} \end{bmatrix}, \tag{12}$$

$$\mathbf{F}^{-1} = \frac{1}{\sqrt{N}} \mathbf{F}^H, \tag{13}$$

$$\lim_{t \to \infty} \frac{||HFC(f^t(\boldsymbol{x}))||_2}{||LFC(f^t(\boldsymbol{x}))||_2} = 0. \tag{14}$$

$$\lim_{t \to \infty} \frac{||HFC(f^t(\boldsymbol{x}))||_2}{||LFC(f^t(\boldsymbol{x}))||_2} = 0. \tag{15}$$

$$\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}, \tag{16}$$

$$f^t(\boldsymbol{x}) = \mathbf{A}^t \boldsymbol{x} = (\mathbf{P}\mathbf{J}\mathbf{P}^{-1})^t \boldsymbol{x}. \tag{17}$$

$$\lim_{t \to \infty} \frac{||HFC[f^t(\boldsymbol{x}) - \lambda_1^t \mathbf{v}_1]||_2}{||LFC[\lambda_1^t \mathbf{v}_1]||_2} = 0 \tag{18}$$

Thanks