Disentangling Past-Future Modeling in Sequential Recommendation via Dual Networks

Hengyu Zhang*
zhang-hy21@mails.tsinghua.edu.cn
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China

Enming Yuan*
yem19@mails.tsinghua.edu.cn
Institute for Interdisciplinary Information Sciences,Tsinghua University
Beijing, China

Wei Guo
guowei67@huawei.com
Huawei Noah's Ark Lab
Shenzhen, China

Zhicheng He
hezhicheng9@huawei.com
Huawei Noah's Ark Lab
Shenzhen, China

Jiarui Qin
qinjr@icloud.com
Shanghai Jiao Tong University
Shanghai, China

Huifeng Guo
huifeng.guo@huawei.com
Huawei Noah's Ark Lab
Shenzhen, China

Bo Chen
chenbo116@huawei.com
Huawei Noah's Ark Lab
Shenzhen, China

Xiu Li†
li.xiu@sz.tsinghua.edu.cn
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China

Ruiming Tang†
tangruiming@huawei.com
Huawei Noah's Ark Lab
Shenzhen, China

CIKM 2022

Reported by Minqin Li

Introduction

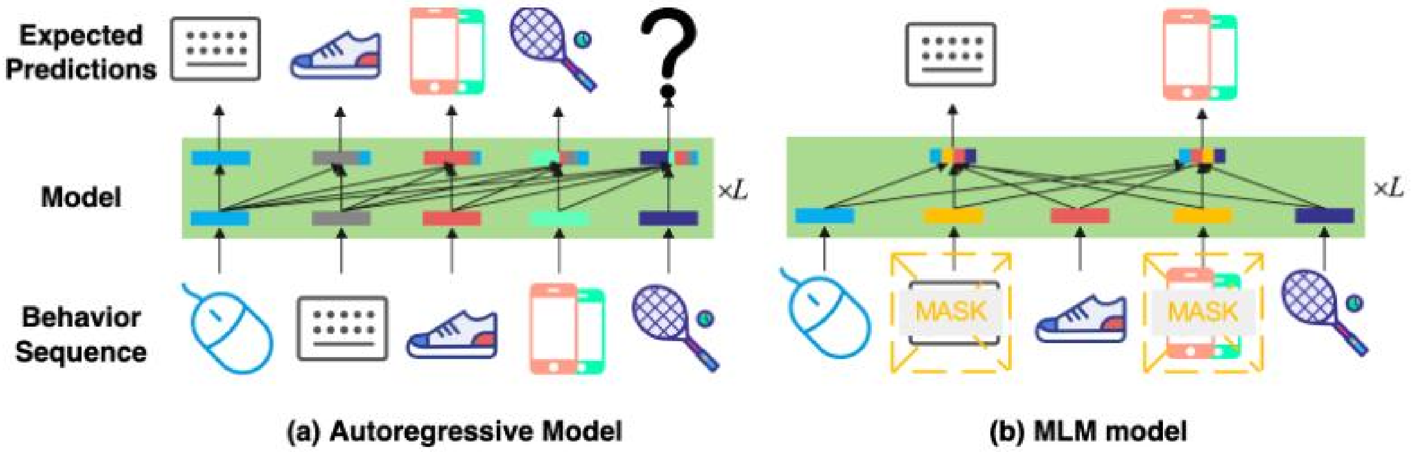


Figure 1: Illustration of common sequential recommendation models.

Method

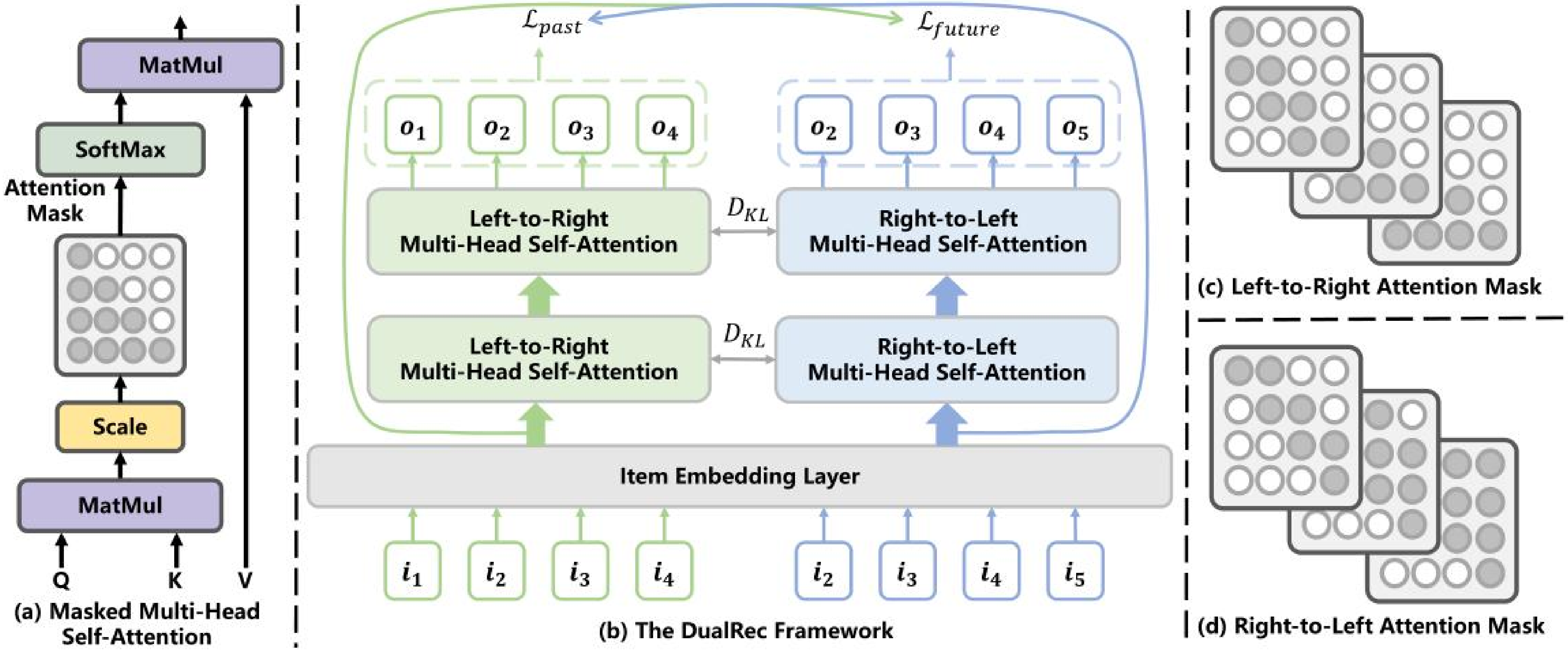
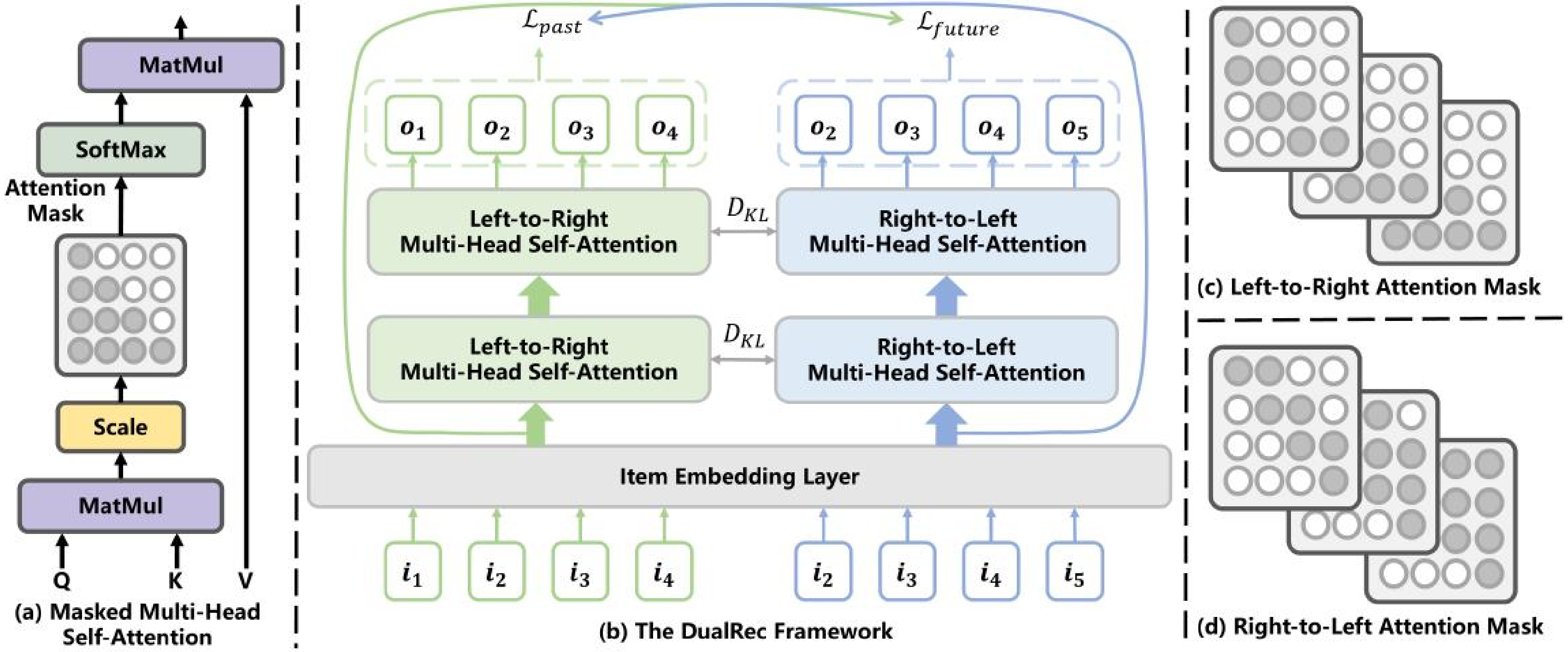


Figure 2: Model architecture of the proposed DualRec framework. (a) Illustration of masked multi-head self-attention. The shaded nodes are visible. (b) Overall structure of the dual network, with the past encoder on the left and the future encoder on the right. Past and future encoders are associated by the shared embedding layer and bi-directional information transferring using KL divergences. (c) and (d) are the attention masks from left to right in the past encoder and from right to left in the future encoder, respectively, and the time windows corresponding to each attention head are different, i.e., multi-scale.

Method



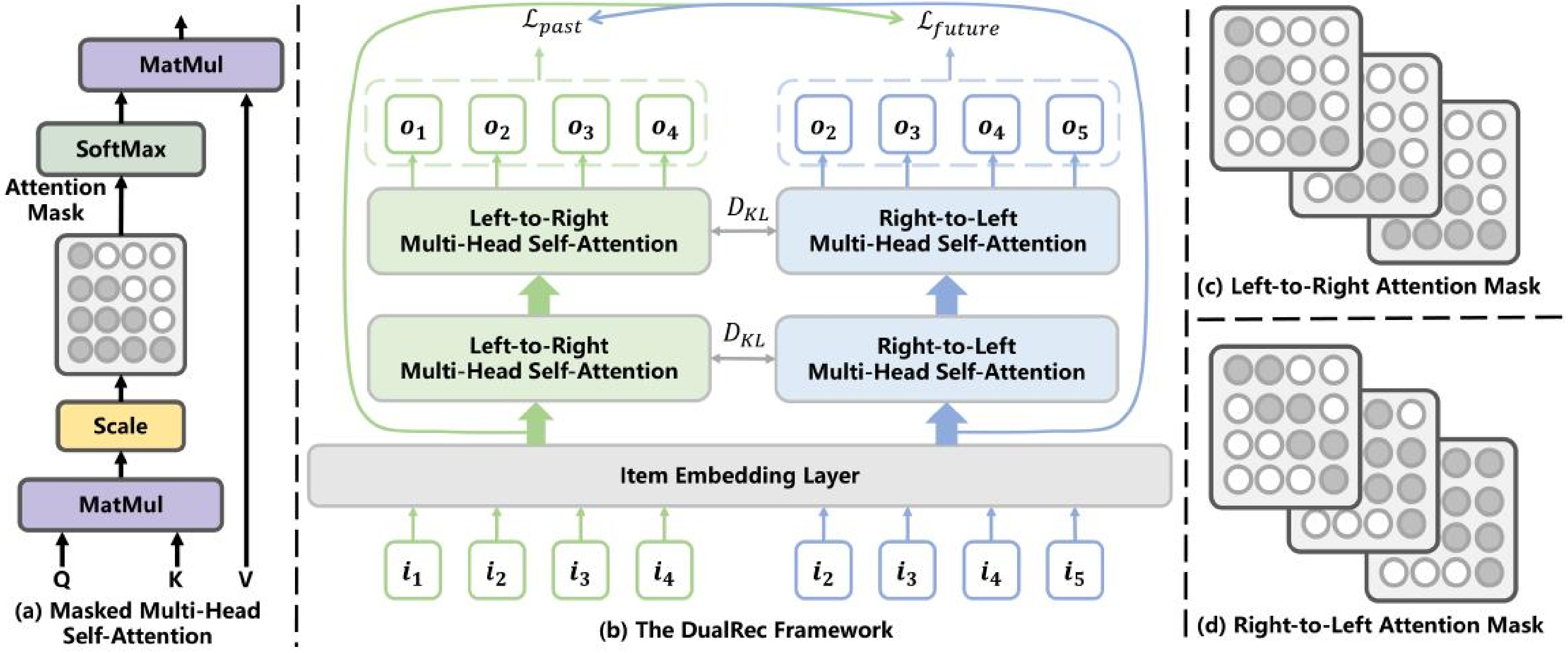
$$p(i_{T_u+1}^{(u)} = i^{(c)} | \mathcal{S}^{(u)}) = \text{SeqRecModel}(\mathcal{S}^{(u)}, i^{(c)}) \quad (1)$$

$$\mathbf{X}^{(0)} = (\mathbf{e}_1, \cdots, \mathbf{e}_n), \ \mathbf{e}_k = \text{LookUp}(i_k, \mathbf{E}^{\mathcal{I}}) \quad (2)$$

$$\mathbf{p}(i, j) = \text{LookUp}(\text{Dist}(i, j), \mathbf{E}^{\mathcal{P}}) \quad (3)$$

$$\mathbf{head}_i = softmax\left(\frac{(\mathbf{X}\mathbf{W}_i^Q) \cdot (\mathbf{X}\mathbf{W}_i^K)^\top}{\sqrt{d/h}}\right)(\mathbf{X}\mathbf{W}_i^V) \quad (4)$$

Method



(a) Masked Multi-Head Self-Attention

(b) The DualRec Framework

(c) Left-to-Right Attention Mask

(d) Right-to-Left Attention Mask

$$\mathrm{MSA}(\mathbf{X}) = \mathrm{Concat}(\mathbf{head}_i, \mathbf{head}_2, \ldots, \mathbf{head}_h)\mathbf{W}^o \quad (5)$$

$$\mathrm{PFF}(\mathbf{X}) = \mathrm{FC}(\sigma(\mathrm{FC}(\mathbf{X}))), \mathrm{FC}(\mathbf{X}) = \mathbf{X}\mathbf{W} + \mathbf{b} \quad (6)$$

$$\begin{aligned} \mathbf{H}^{(l)} &= \mathrm{LayerNorm}\left(\mathbf{X}^{(l-1)} + \mathrm{MSA}(\mathbf{X}^{(l-1)})\right), \\ \mathbf{X}^{(l)} &= \mathrm{LayerNorm}\left(\mathbf{H}^{(l)} + \mathrm{PFF}(\mathbf{H}^{(l)})\right), \end{aligned} \quad (7)$$

Method

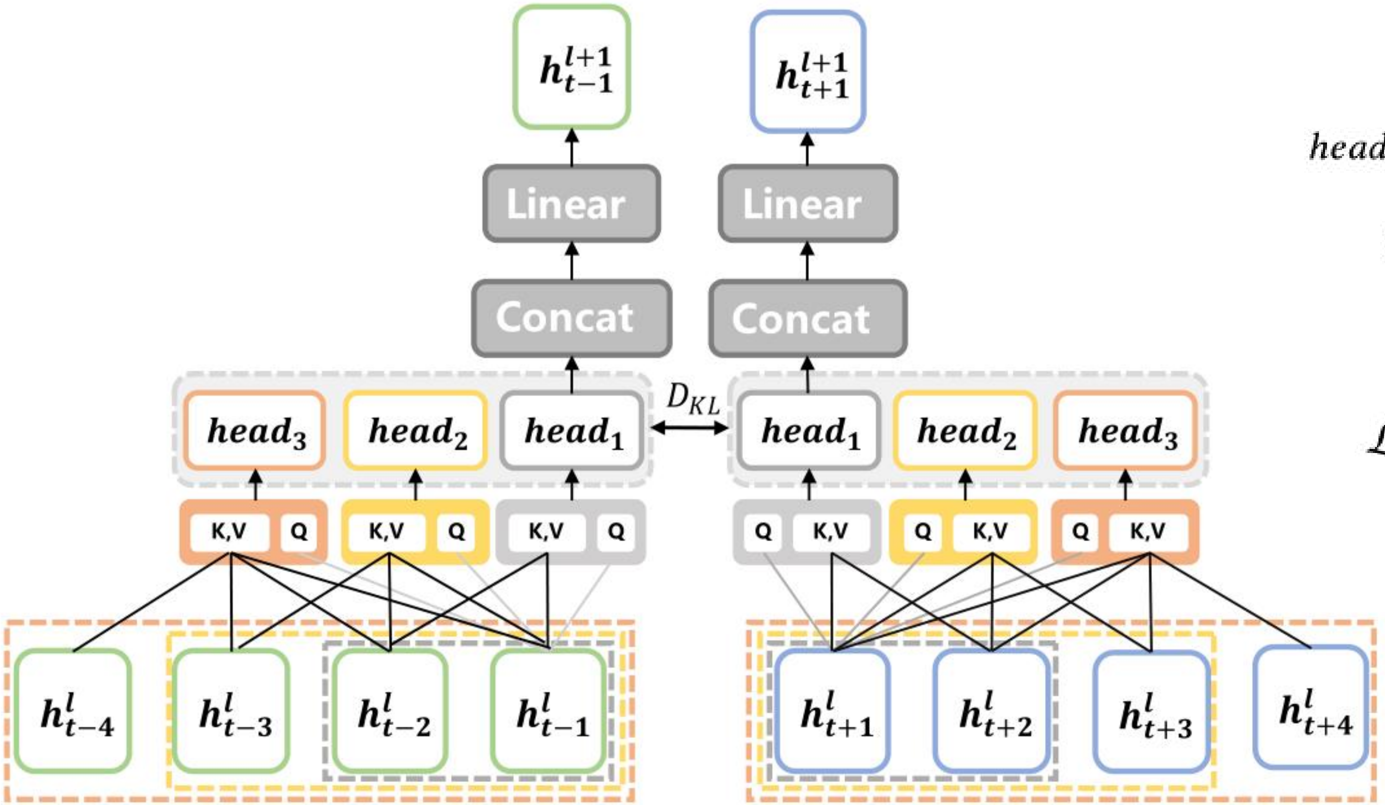


Figure 3: Illustration of interest-level knowledge transferring between past and future information.

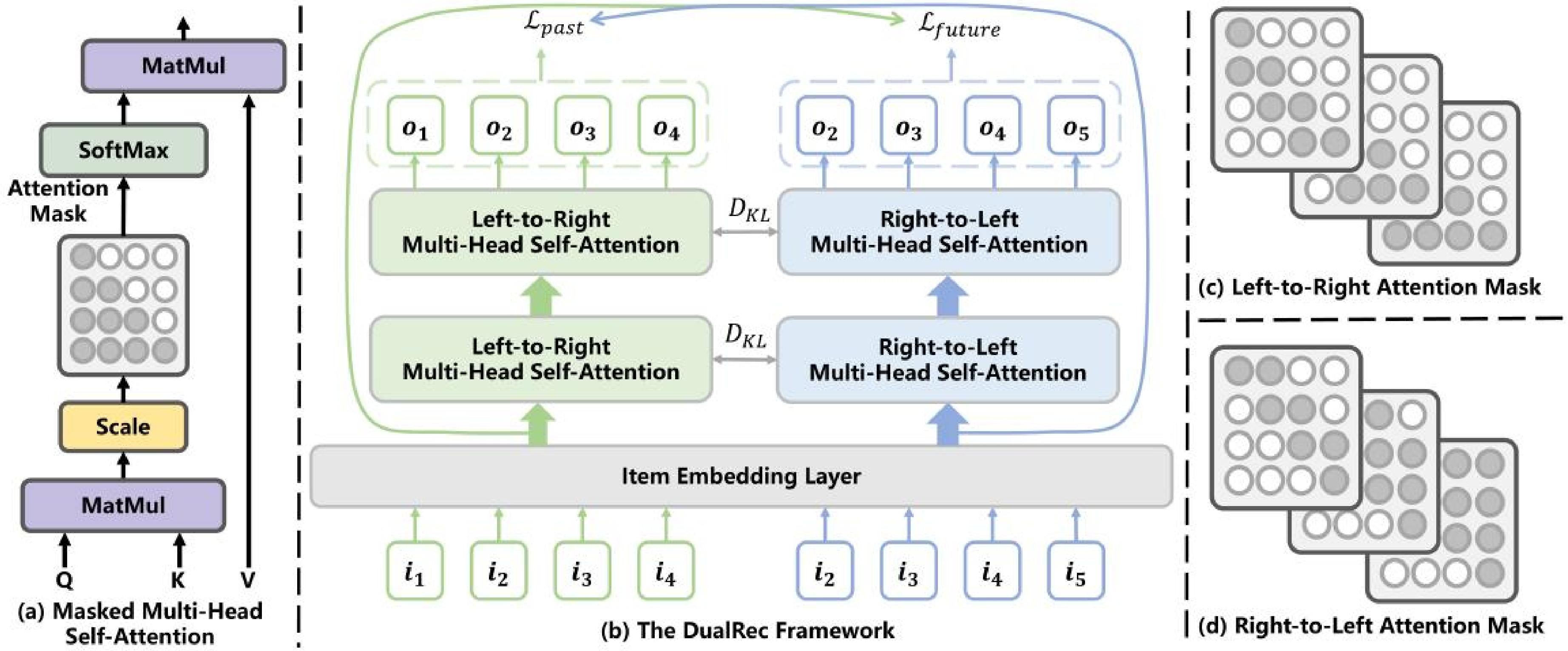
$$WS(i) = \begin{cases} i+1 & \text{if } i \le \frac{h}{2} \\ \frac{h}{2} + \left[\frac{\exp(i-\frac{h}{2})}{\exp \frac{h}{2}} \cdot \left(n - \frac{h}{2}\right) \right] & \text{if } i > \frac{h}{2} \end{cases}, \quad i = 1 \cdots h \tag{8}$$

$$\begin{aligned} head_{i,j}(\mathbf{X}, \sigma) &= softmax\left(\frac{(\mathbf{X}\mathbf{W}_i^Q)_j \cdot S_j\left(\mathbf{X}\mathbf{W}_i^K, \sigma\right)^\top}{\sqrt{d/h}} \right) \cdot S_j\left(\mathbf{X}\mathbf{W}_i^V, \sigma\right), \\ S_j(\mathbf{X}, \sigma) &= \left[\mathbf{X}_{j-\sigma}, \cdots, \mathbf{X}_j\right], \end{aligned} \tag{9}$$

$$\mathcal{L}_{reg} = \sum_{i=1}^{h} \frac{1}{2} \left(D_{KL}(\mathbf{head}_i^p || \mathbf{head}_i^f) + D_{KL}(\mathbf{head}_i^f || \mathbf{head}_i^p) \right), \tag{10}$$

$$D_{KL}(p||q) = \sum_i p_i \log \frac{p_i}{q_i}, \tag{11}$$

Method



$$\mathbf{s}_t^p = \mathbf{o}_{t-1}^p \cdot \mathbf{E}^{\mathcal{I}\top}, \quad \mathbf{s}_t^f = \mathbf{o}_{t+1}^f \cdot \mathbf{E}^{\mathcal{I}\top} \tag{12}$$

$$\hat{\mathbf{y}}_t^p = softmax(\mathbf{s}_t^p), \quad \hat{\mathbf{y}}_t^f = softmax(\mathbf{s}_t^f) \tag{13}$$

$$\begin{aligned}\mathcal{L} &= \alpha\,\mathcal{L}_{past} + (1-\alpha)\,\mathcal{L}_{future} + \beta\,\mathcal{L}_{reg} \\ &= -\alpha \sum_{t=2}^{n-1} \mathrm{OneHot}(i_t^\star)\log \hat{\mathbf{y}}_t^p - (1-\alpha)\sum_{t=2}^{n-1} \mathrm{OneHot}(i_t^\star)\log \hat{\mathbf{y}}_t^f + \beta\,\mathcal{L}_{reg},\end{aligned} \tag{14}$$

Experiments

Table 1: Dataset Statistics

Dataset	# Users	# Items	# Interactions	Density
Beauty	22,363	12,101	198,502	0.07%
Sports	25,598	18,357	296,337	0.05%
Toys	19,412	11,924	167,597	0.07%
Yelp	30,431	20,033	316,354	0.05%

Experiments

Table 2: Performance comparison using different methods on four popular sequential datasets. The best performance and the second-best performance methods are denoted in bold and underlined fonts respectively. The "*" mark denotes the statistical significance ($p < 0.05$) of comparing DualRec with the strongest baseline results and the "Improv." column represents the relative improvement of DualRec over the strongest baseline.

Datasets	Metric	GRU4Rec	Caser	HGN	RepeatNet	CLEA	SASRec	S3-Rec$_{MIP}$	BERT4Rec	SRGNN	GCSAN	FMLP-Rec	DualRec	Improv.
Beauty	HR@1	0.1519	0.1337	0.1683	0.1578	0.1325	0.1907	0.1678	0.1531	0.1729	0.1973	0.2051	**0.2289***	+11.60%
	HR@5	0.3612	0.3032	0.3544	0.3268	0.3305	0.4036	0.3710	0.3640	0.3518	0.3678	0.4103	**0.4241***	+3.24%
	NDCG5	0.2608	0.2219	0.2656	0.2455	0.2353	0.3022	0.2735	0.2622	0.2660	0.2864	0.3133	**0.3320***	+5.97%
	HR@10	0.4657	0.3942	0.4503	0.4205	0.4426	0.5043	0.4749	0.4739	0.4484	0.4542	0.5070	**0.5190***	+2.37%
	NDCG@10	0.2944	0.2512	0.2965	0.2757	0.2715	0.3358	0.3069	0.2975	0.2971	0.3143	0.3443	**0.3626***	+5.32%
	MRR	0.2593	0.2263	0.2669	0.2498	0.2376	0.2990	0.2731	0.2614	0.2686	0.2882	0.3102	**0.3302***	+6.45%
Sports	HR@1	0.1366	0.1135	0.1428	0.1334	0.1114	0.1676	0.1107	0.1255	0.1419	0.1669	0.1722	**0.1947***	+13.07%
	HR@5	0.3552	0.2866	0.3349	0.3162	0.3041	0.3919	0.3141	0.3375	0.3367	0.3588	0.3886	**0.4127***	+5.31%
	NDCG5	0.2487	0.2020	0.2420	0.2274	0.2096	0.2823	0.2143	0.2341	0.2418	0.2658	0.2839	**0.3080***	+8.49%
	HR@10	0.4853	0.4014	0.4551	0.4324	0.4274	0.5169	0.4491	0.4772	0.4545	0.4737	0.5098	**0.5383***	+4.14%
	NDCG@10	0.2907	0.2390	0.2806	0.2649	0.2493	0.3244	0.2578	0.2775	0.2799	0.3029	0.3231	**0.3485***	+7.43%
	MRR	0.2493	0.2100	0.2469	0.2334	0.2156	0.2838	0.2203	0.2378	0.2461	0.2691	0.2830	**0.3078***	+8.47%
Toys	HR@1	0.1303	0.1114	0.1504	0.1333	0.1104	0.1760	0.1825	0.1262	0.1600	0.1996	0.2003	**0.2268***	+13.23%
	HR@5	0.3526	0.2614	0.3276	0.3001	0.3055	0.3975	0.3892	0.3344	0.3389	0.3613	0.4010	**0.4152***	+3.54%
	NDCG5	0.2444	0.1885	0.2423	0.2192	0.2102	0.2907	0.2903	0.2327	0.2528	0.2836	0.3055	**0.3253***	+6.48%
	HR@10	0.4691	0.3540	0.4211	0.4015	0.4207	0.5034	0.4935	0.4493	0.4413	0.4509	0.4977	**0.5145***	+2.21%
	NDCG@10	0.2820	0.2183	0.2724	0.2517	0.2473	0.3271	0.3239	0.2698	0.2857	0.3125	0.3367	**0.3573***	+6.12%
	MRR	0.2424	0.1967	0.2454	0.2253	0.2138	0.2877	0.2890	0.2338	0.2566	0.2871	0.3034	**0.3256***	+7.32%
Yelp	HR@1	0.1970	0.2188	0.2428	0.2341	0.2102	0.2327	0.2250	0.2405	0.2176	0.2493	0.2625	**0.2893***	+10.21%
	HR@5	0.5788	0.5111	0.5768	0.5357	0.5707	0.5949	0.5978	0.5976	0.5442	0.5725	0.6246	**0.6328***	+1.32%
	NDCG5	0.3933	0.3696	0.4162	0.3894	0.3955	0.4198	0.4171	0.4252	0.3860	0.4162	0.4507	**0.4681***	+3.93%
	HR@10	0.5788	0.6661	0.7411	0.6897	0.7473	0.7722	0.7764	0.7597	0.7096	0.7371	0.7699	**0.7896***	+2.25%
	NDCG@10	0.4511	0.4198	0.4695	0.4393	0.4527	0.4790	0.4751	0.4778	0.4395	0.4696	0.4981	**0.5190***	+4.20%
	MRR	0.3684	0.3595	0.3998	0.3769	0.3751	0.3994	0.3938	0.4026	0.3711	0.4006	0.4236	**0.4466***	+5.43%

Experiments

Table 3: Ablation study results of different consisting components in DualRec.

Methods	Beauty		Sports		Toys		Yelp	
	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5	HR@5	NDCG@5
DualRec	**0.4241**	**0.3320**	**0.4127**	**0.3080**	**0.4152**	**0.3253**	**0.6328**	**0.4681**
DualRec w/o DE	0.4154	0.3251	0.4037	0.3016	0.4098	0.3202	0.6273	0.4610
DualRec w/o BIT	0.4235	0.3311	0.4102	0.3055	0.4122	0.3246	0.6284	0.4658
DualRec w/o RPE	0.4216	0.3306	0.4093	0.3044	0.413	0.3234	0.6320	0.4677
DualRec w/o LN	0.4086	0.3075	0.4097	0.3031	0.3987	0.2963	0.6321	0.4653
DualRec w/o RC	0.3653	0.2642	0.3720	0.2632	0.3388	0.2360	0.5751	0.4052

Experiments

Table 4: Compatibility Analysis with Different Backbones on four sequential datasets. Prefix "Dual" indicates the corresponding dual network model of backbone, and "BIT" means bi-directional information transferring.

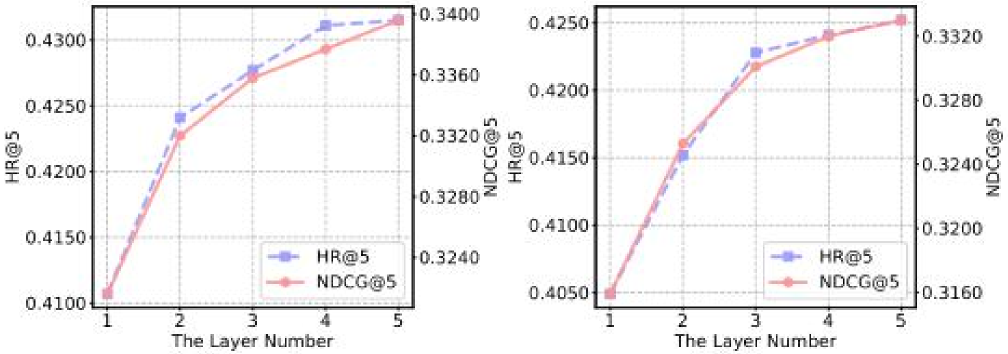
Datasets	Beauty			Sports			Toys			Yelp		
Model	HR@5	NDCG@5	MRR	HR@5	NDCG@5	MRR	HR@5	NDCG@5	MRR	HR@5	NDCG@5	MRR
SASRec	0.4036	0.3022	0.2990	0.3919	0.2823	0.2838	0.3975	0.2907	0.2877	0.5949	0.4198	0.3994
DualSASRec	0.4178	0.3147	0.3108	0.4055	0.2946	0.2936	0.4068	0.3038	0.3009	0.6198	0.4415	0.4183
DualSASRec+BIT	**0.4223**	**0.3199**	**0.3153**	**0.4111**	**0.2985**	**0.2961**	**0.4111**	**0.3084**	**0.3050**	**0.6199**	**0.4457**	**0.4239**
GRU4Rec	0.3612	0.2608	0.2593	0.3552	0.2487	0.2493	0.3526	0.2444	0.2424	0.5788	0.3933	0.3684
DualGRU4Rec	0.3779	0.2726	0.2696	0.3681	0.2578	0.2575	0.3532	0.2454	0.2443	0.5869	0.4036	0.3803
DualGRU4Rec+BIT	**0.3866**	**0.2826**	**0.2798**	**0.3768**	**0.2638**	**0.2624**	**0.3679**	**0.2567**	**0.2535**	**0.5988**	**0.4141**	**0.3888**
FMLP-Rec	0.4103	0.3133	0.3102	0.3886	0.2839	0.2830	0.4010	0.3055	0.3034	0.6246	0.4507	0.4236
DualFMLP-Rec	0.4172	0.3200	0.3170	0.3937	0.2889	0.2878	0.4041	0.3084	0.3063	**0.6351**	**0.4708**	**0.4454**
DualFMLP-Rec+BIT	**0.4210**	**0.3240**	**0.3208**	**0.4002**	**0.2940**	**0.2914**	**0.4067**	**0.3109**	**0.3087**	0.6343	0.4668	0.4407

Experiments

Table 5: The performance of Past-Only Model, Future-Only Model, and DualRec on four datasets.

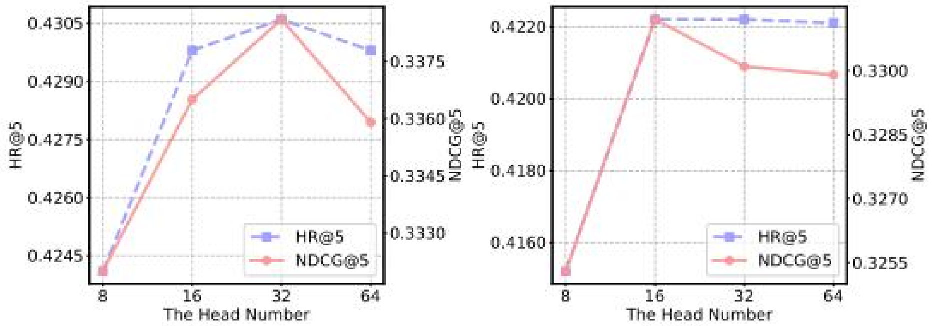
Datasets	Metric	Past-Only	Future-Only	DualRec
Beauty	HR@5	0.4166	0.3833	**0.4235**
	NDCG@5	0.3272	0.2912	**0.3311**
Sports	HR@5	0.4061	0.3786	**0.4102**
	NDCG@5	0.3027	0.2766	**0.3055**
Toys	HR@5	0.4085	0.3601	**0.4122**
	NDCG@5	0.3212	0.2761	**0.3246**
Yelp	HR@5	0.6276	0.6215	**0.6284**
	NDCG@5	0.4628	0.4576	**0.4658**

Experiments



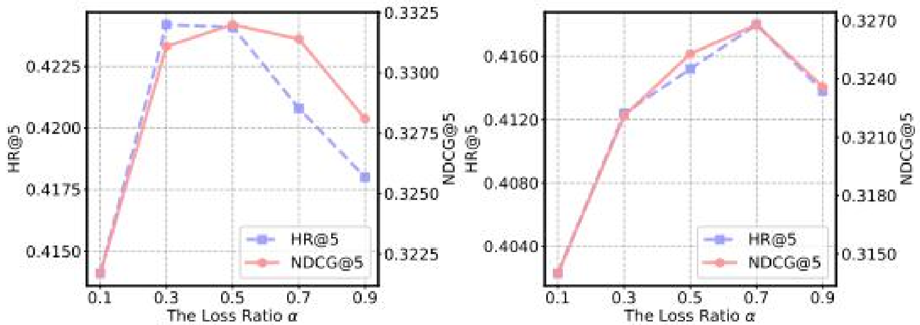
(a) Beauty (b) Toys

Figure 4: Performances with different layer number.



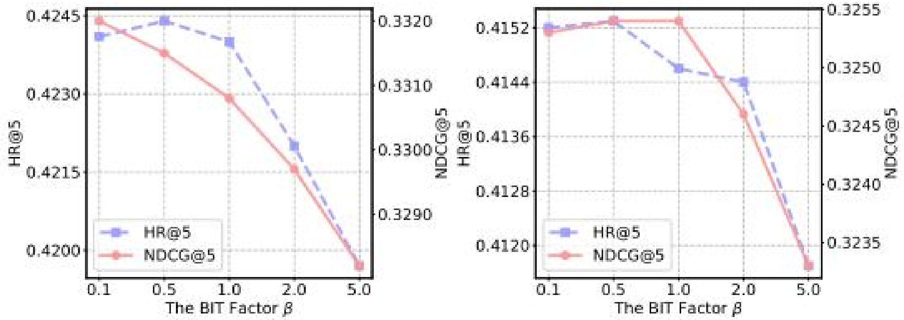
(a) Beauty (b) Toys

Figure 5: Performances with different attention head number.



(a) Beauty (b) Toys

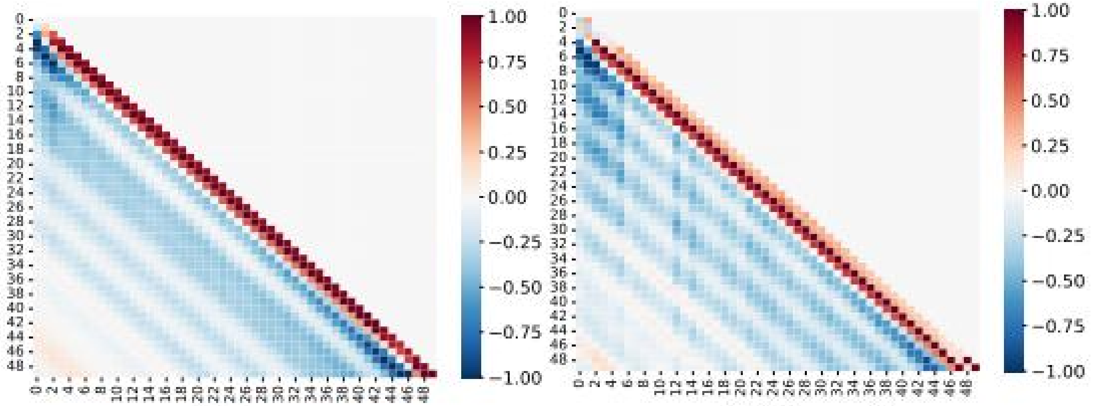
Figure 6: Performances with different loss ratio of dual network model.



(a) Beauty (b) Toys

Figure 7: Performances with different factor of bi-directional information transferring loss.

Experiments



(a) Layer1 (b) Layer2

Figure 8: Heat maps of the average difference in attention score per layer between the past encoder in the DualRec and the Past-Only Model on the Yelp test dataset (red means the average attention scores in the DualRec are higher than those in the Past-Only Model; and blue indicates the opposite case). The coordinates in the figure indicate the sequence index.

Thanks