Recommender Transformers with Behavior Pathways

Zhiyu Yao[†], Xinyang Chen[†], Sinan Wang[‡], Qinyan Dai[*],
Yumeng Li[‡], Tanchao Zhu[‡], Mingsheng Long[†](✉)
School of Software, BNRist, Tsinghua University, China[†]
Alibaba Group, China[‡]
Industrial Engineering, Tsinghua University[*]
{yaozy19,chenxiny17,dai-qy18}@mails.tsinghua.edu.cn
mingsheng@mails.tsinghua.edu.cn

code:none

WWW 2024

Reported by Minqin Li

Introduction

Figure 1: Three main characteristics of the behavior pathway for different users, making sequential recommendation extremely hard. The behavior pathway is outlined by the red boxes.

In sequential recommendation, we find that only a small part of pivotal behaviors can be evolved into the use's future action. And we conclude this characteristics of sequential behaviors as the Behavior Pathway.

Method

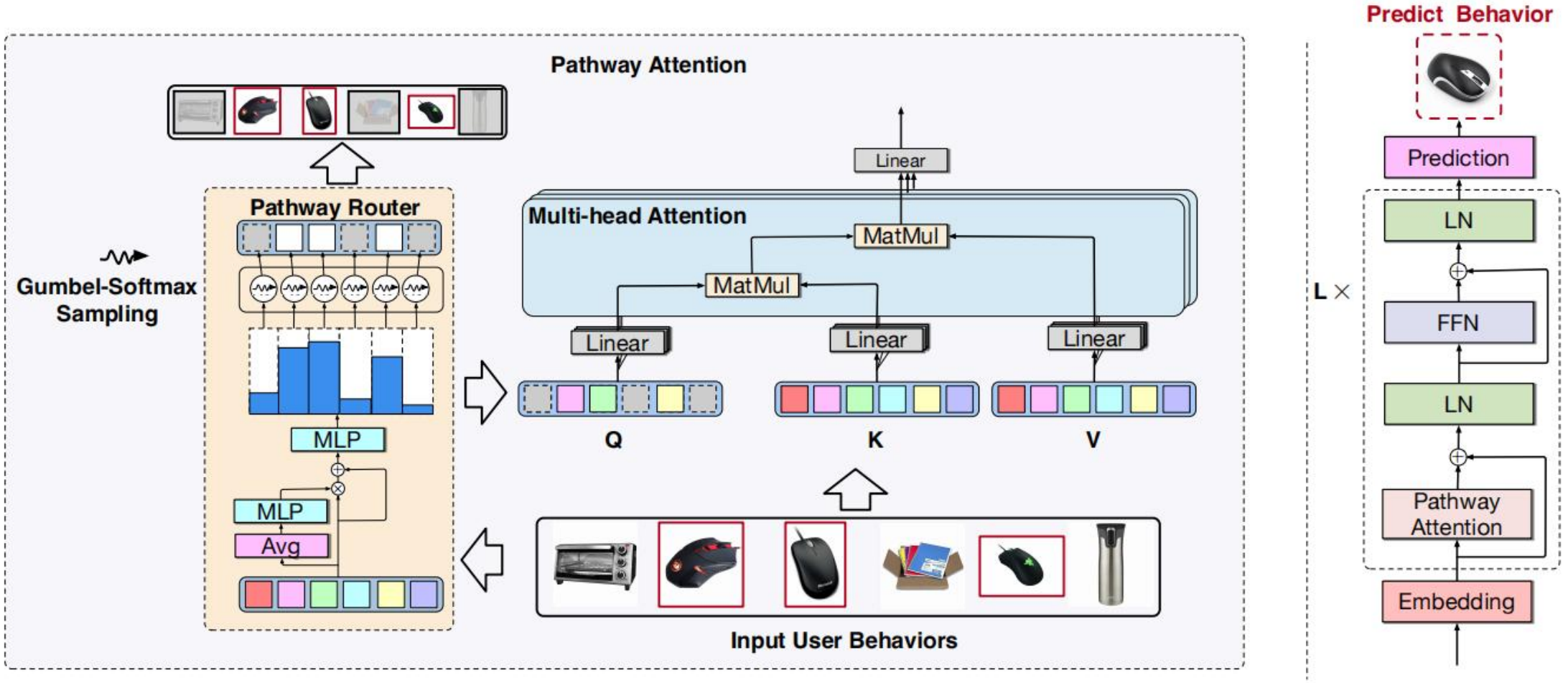
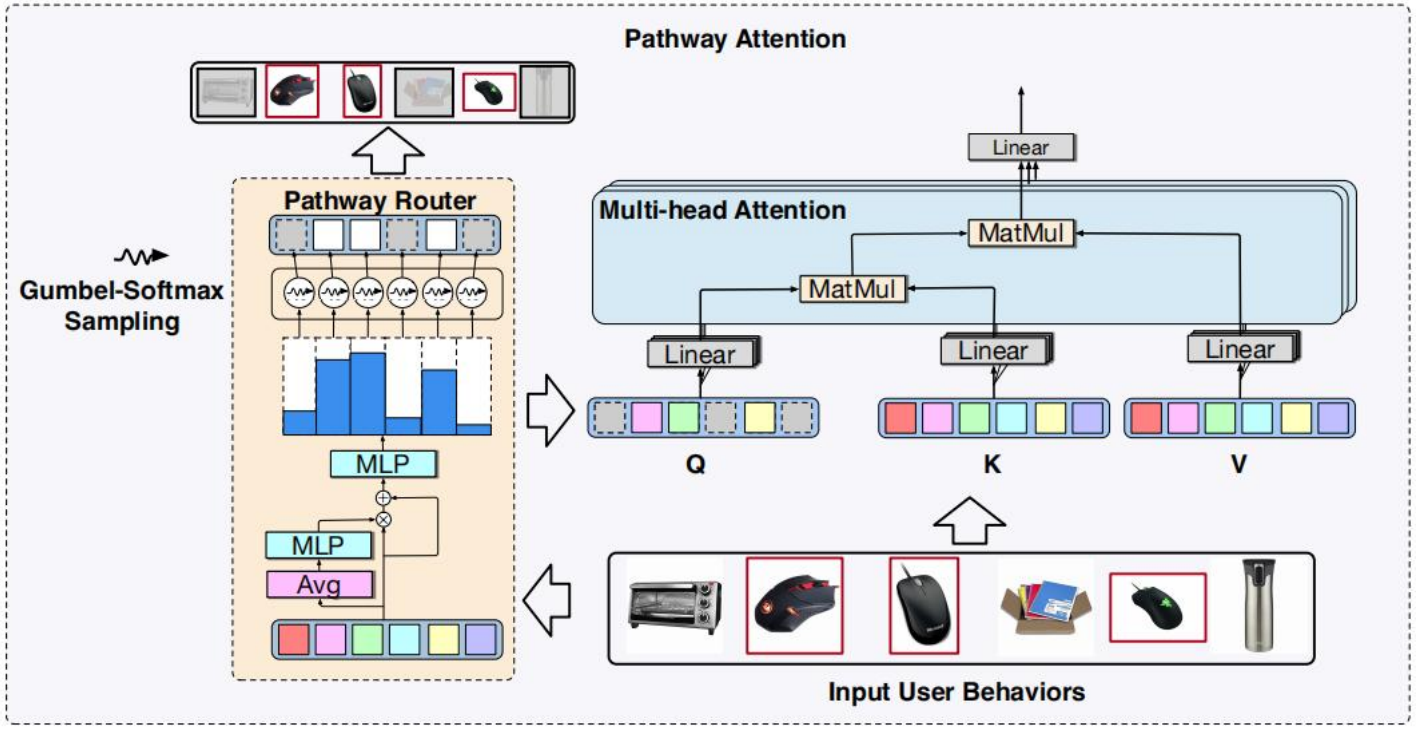


Figure 2: Recommender Transformer architecture (right). Pathway Attention (left) explores the behavior pathway by the pathway router (orange module) and captures the evolving sequential characteristics by the multi-head attention.

Method



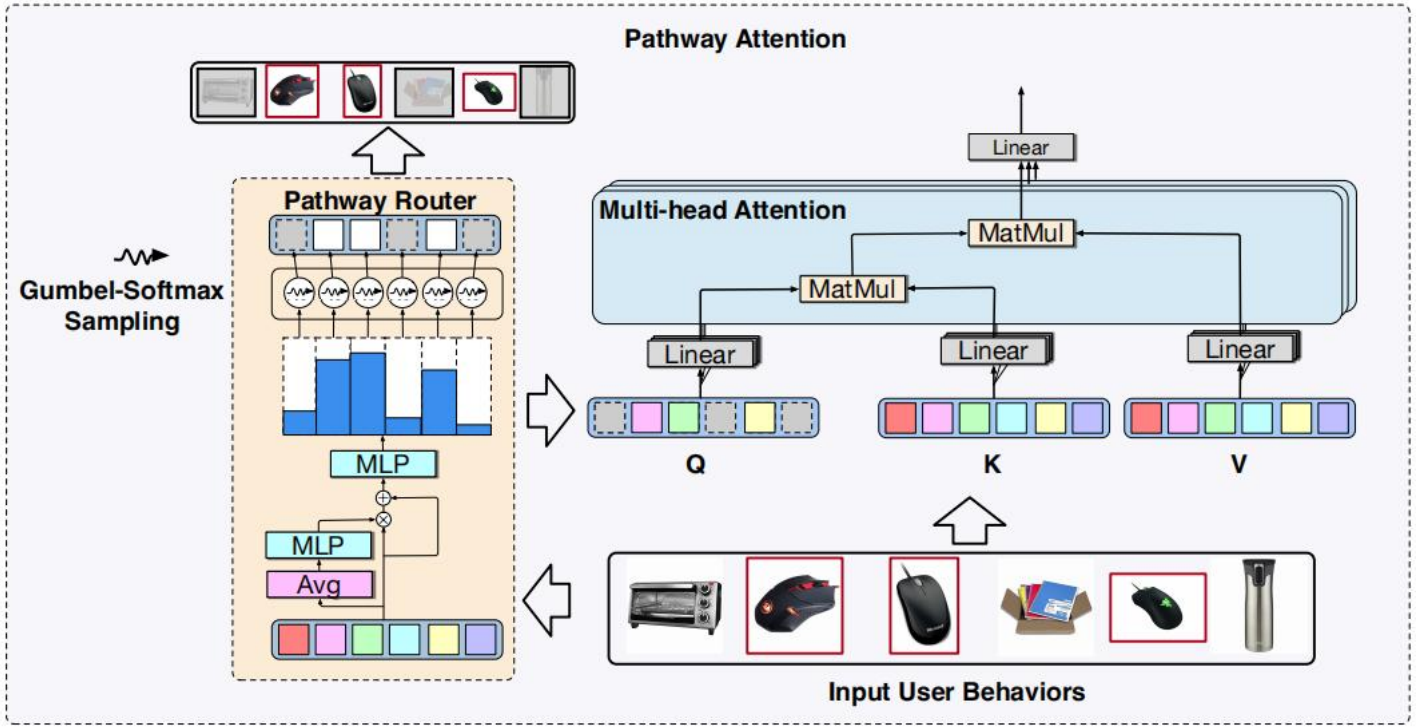
$$\widehat{\mathcal{Z}}^l, \mathcal{R}^l = \text{Path-MSA}\left(\mathcal{Z}^{l-1}, \mathcal{R}^{l-1}\right)$$
$$\widehat{\mathcal{Z}}^l = \text{LN}\left(\widehat{\mathcal{Z}}^l + \mathcal{Z}^{l-1}\right)$$
$$\mathcal{Z}^l = \text{LN}\left(\text{FFN}\left(\widehat{\mathcal{Z}}^l\right) + \widehat{\mathcal{Z}}^l\right), \quad (1)$$

$$\mathcal{Z}^l_{\text{emb}} = \mathcal{Z}^{l-1} + \mathcal{Z}^{l-1} \odot \text{MLP}\left(\frac{\sum_{i=1}^{N} \mathcal{R}_i^{l-1} \mathcal{Z}_i^{l-1}}{\sum_{i=1}^{N} \mathcal{R}_i^{l-1}}\right) \quad (2)$$

$$\boldsymbol{\pi} = \text{Softmax}\left(\text{MLP}(\mathcal{Z}^l_{\text{emb}})\right) \in \mathbb{R}^{N \times 2},$$

$$\widehat{\mathcal{R}}^l_t = \underset{j \in \{0,1\}}{\arg\max} \left(\log \pi_t(j) + G_t(j)\right), \quad (3)$$

Method



$$v_t(j) = \frac{\exp((\log \pi_t(j) + G_t(j))/\tau)}{\sum_{i\in\{0,1\}} \exp((\log \pi_t(i) + G_t(i))/\tau)}, j \in \{0,1\}, \quad (4)$$

$$\mathcal{R}^l = \hat{\mathcal{R}}^l \odot \mathcal{R}^{l-1}. \quad (5)$$

$$\mathcal{Q}_m, \mathcal{K}_m, \mathcal{V}_m = (\mathcal{Z}^{l-1} \odot \mathcal{R}^l) W^l_{\mathcal{Q}_m}, \mathcal{Z}^{l-1} W^l_{\mathcal{K}_m}, \mathcal{Z}^{l-1} W^l_{\mathcal{V}_m}$$

$$\hat{\mathcal{Z}}^l_m = \text{Softmax}\left(\frac{\mathcal{Q}_m \mathcal{K}^{\mathrm{T}}_m}{\sqrt{d/h}}\right) \mathcal{V}^l_m, \quad (6)$$

$$\mathcal{L} = -\sum_{u\in\mathcal{U}} \sum_{t=1}^{n} \log \sigma(p(i_{t+1} \mid i_{1:t}) - p(i^-_{t+1} \mid i_{1:t})), \quad (7)$$

Experiments

Table 1: Statistics of the datasets.

Dataset	Users	Items	Actions
Beauty	22,363	12,101	19,8502
Sports	25,598	18,357	29,6337
Toys	19,412	11,924	16,7597
Yelp	30,431	20,033	31,6354
MovieLens-1M	6,040	3,416	1,000,000
Tmall	66,909	37,367	42,7797
Steam	334,730	13,047	3,700,000

Experiments

Datasets	Meric	PopRec	Caser	GRU4Rec	BERT4Rec	SASRec	SASRec+	SMRec	RETR
Beauty	HR@10	0.3386	0.3942	0.4106	0.4739	0.4696	0.4798	0.4826	**0.5034**
	NDCG@10	0.1803	0.2512	0.2584	0.2975	0.3156	0.3261	0.3238	**0.3425**
	MRR	0.1558	0.2263	0.2308	0.2614	0.2852	0.2901	0.2918	**0.3067**
Sports	HR@10	0.3423	0.4014	0.4299	0.4722	0.4622	0.4776	0.4853	**0.5083**
	NDCG@10	0.1902	0.2390	0.2527	0.2775	0.2869	0.2987	0.3061	**0.3175**
	MRR	0.1660	0.2100	0.2191	0.2378	0.2520	0.2635	0.2665	**0.2768**
Toys	HR@10	0.3008	0.3540	0.3896	0.4493	0.4663	0.4729	0.4754	**0.5104**
	NDCG@10	0.1618	0.2183	0.2274	0.2698	0.3136	0.3183	0.3198	**0.3395**
	MRR	0.1430	0.1967	0.1973	0.2338	0.2842	0.2912	0.2910	**0.3048**
Yelp	HR@10	0.3609	0.6661	0.7265	0.7597	0.7373	0.7481	0.7548	**0.7730**
	NDCG@10	0.2007	0.4198	0.4375	0.4778	0.4642	0.4757	0.4789	**0.5136**
	MRR	0.1740	0.3595	0.3630	0.4026	0.3927	0.4011	0.4023	**0.4354**
MovieLen	HR@10	0.4329	0.7886	0.5581	0.8269	0.8233	0.8291	0.8302	**0.8467**
	NDCG@10	0.2377	0.5538	0.3381	0.5965	0.5936	0.6057	0.6079	**0.6351**
	MRR	0.1891	0.5178	0.3002	0.5614	0.5573	0.5649	0.5703	**0.5921**
Tmall	HR@10	0.2967	0.5943	0.6432	0.6196	0.6275	0.6392	0.6476	**0.7138**
	NDCG@10	0.1874	0.4513	0.5169	0.5025	0.5049	0.5169	0.5192	**0.6103**
	MRR	0.1723	0.4209	0.4975	0.4026	0.4804	0.4912	0.4934	**0.5822**
Steam	HR@10	0.7172	0.7874	0.4190	0.8656	0.8729	0.8773	0.8792	**0.9001**
	NDCG@10	0.4535	0.5381	0.2691	0.6283	0.6306	0.6397	0.6408	**0.6795**
	MRR	0.4102	0.5091	0.2402	0.5883	0.5925	0.6005	0.6011	**0.6326**

Experiments

Table 3: Ablation study of (**Left**) the effectiveness of each model component and (**Right**) the number of blocks for each RETR block. Experiments are conducted on the Beauty Dataset.

Model	MRR	Model (# number of blocks)	MRR
RETR	**0.3067**	RETR ($L = 1$)	0.2966
RETR w/o Pathway Router	0.2793	RETR ($L = 2$)	**0.3067**
RETR w/o hierarchical update	0.2957	RETR ($L = 3$)	0.3058
SASRec	0.2852	**RETR** ($L = 4$)	0.3052

Table 4: Ablation study of (**Left**) the effectiveness of different temperatures; Comparison Parameters and GFLOPs (**Right**). All ablation study experiments are conducted on the Yelp Dataset.

Model (temperature)	MRR	Model	Parameters (M)	GFLOPs	MRR
RETR ($\tau = 0.4$)	0.4312	RETR	5.021	9.558	**0.4354**
RETR ($\tau = 0.8$)	**0.4354**	SASRec [16]	4.916	9.552	0.3927
RETR ($\tau = 1$)	0.4292	SASRec+ [35]	5.133	9.942	0.4011
RETR ($\tau = 2$)	0.4183	SMRec [6]	5.173	9.864	0.4023

Experiments

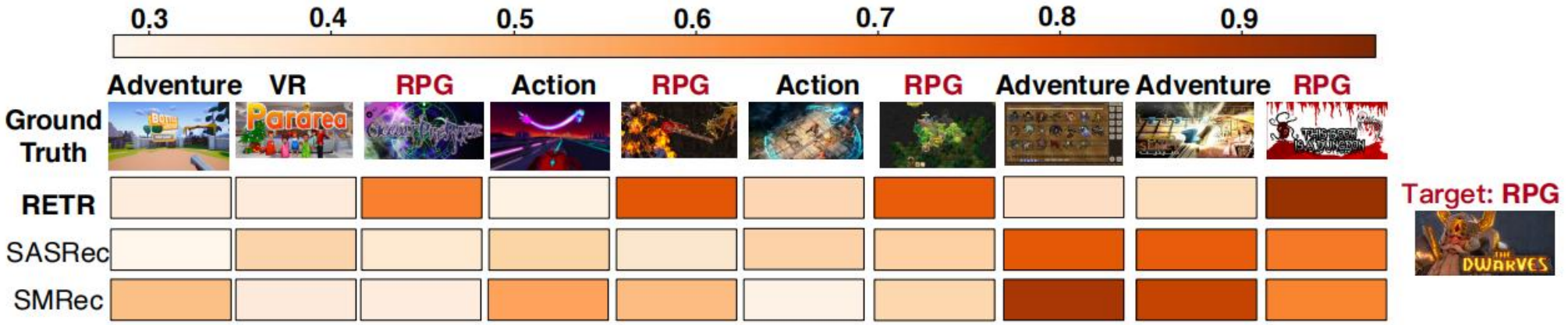


Figure 3: Illustration of how RETR, SASRec and SMRec differs on utilizing the historical behaviors of a random User in Steam Dataset. We provide the visualization of behavior heat maps for RETR, SASRec and SMRec of a random user in Steam dataset.

Thanks