Sparse Attentive Memory Network for Click-through Rate Prediction with Long Sequences

Qianying Lin
Alibaba Group
Hangzhou, China
qianying.lqy@alibaba-inc.com

Wen-Ji Zhou
Alibaba Group
Hangzhou, China
eric.zwj@alibaba-inc.com

Yanshi Wang
Alibaba Group
Hangzhou, China
yanshi.wys@alibaba-inc.com

Qing Da
Alibaba Group
Hangzhou, China
daqing.dq@alibaba-inc.com

Qing-Guo Chen
Alibaba Group
Hangzhou, China
qingguo.cqg@alibaba-inc.com

Bing Wang
Alibaba Group
Hangzhou, China
lingfeng.wb@alibaba-inc.com

CIKM 2022

Code: https://github.com/waldenlqy/SAM

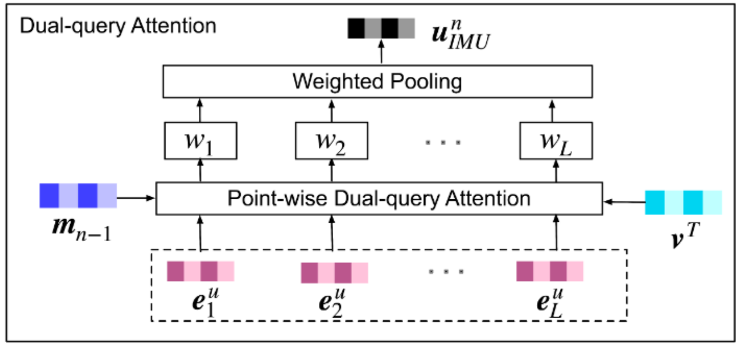
Reported by liang li

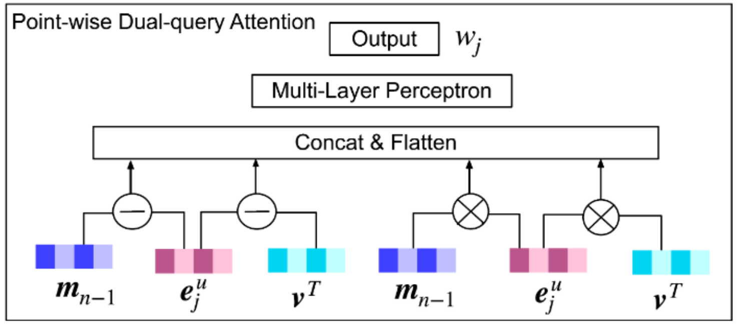
Motivation

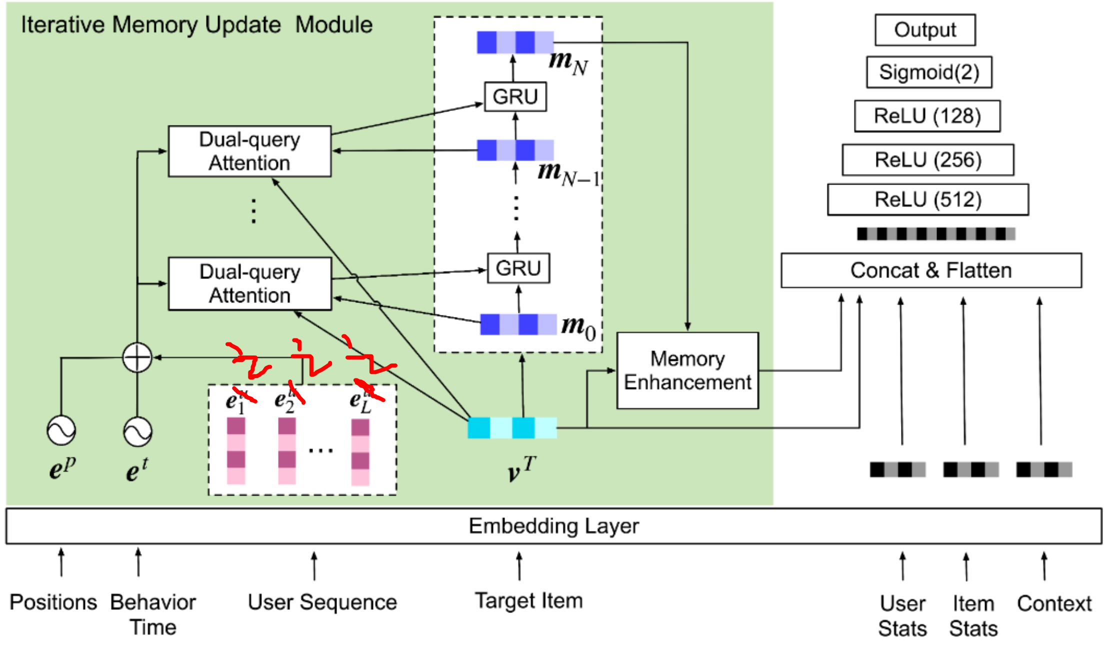
Details：

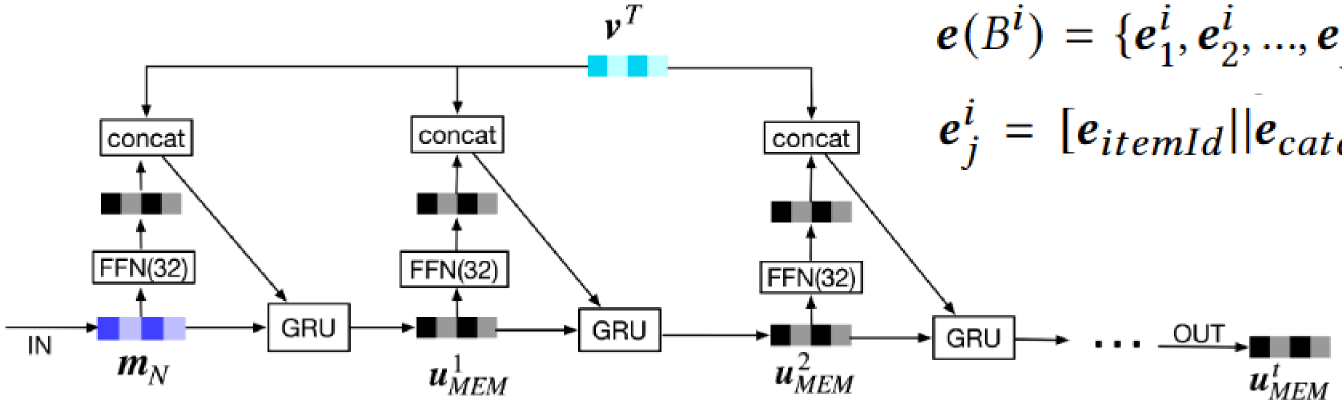
- The introduction of long-term interests improves both recommendation accuracy and the degree of personalization. The sequences used are usually truncated to users’ most recent 50 to 100 behaviors.

Problem Statement









$$e(B^i) = \{e^i_1, e^i_2, ..., e^i_L\},\ e^i_j \in \mathbb{R}^{d_i}$$

$$e^i_j = [e_{itemId}||e_{cateId}||e_{shopId}||e_{brandId}]$$

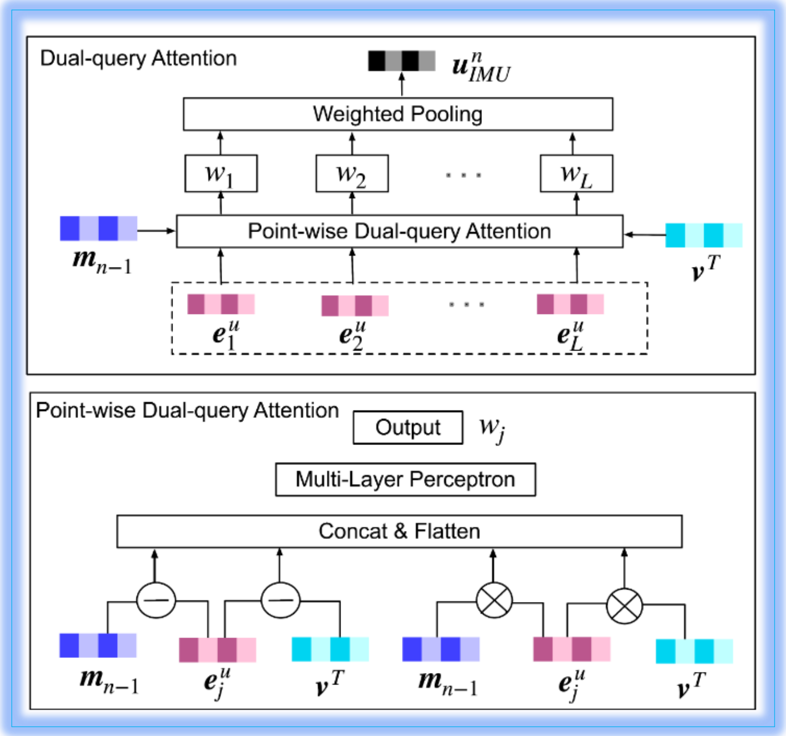
$$e(B^p) = \{e^p_1, e^p_2, ..., e^p_L\}$$

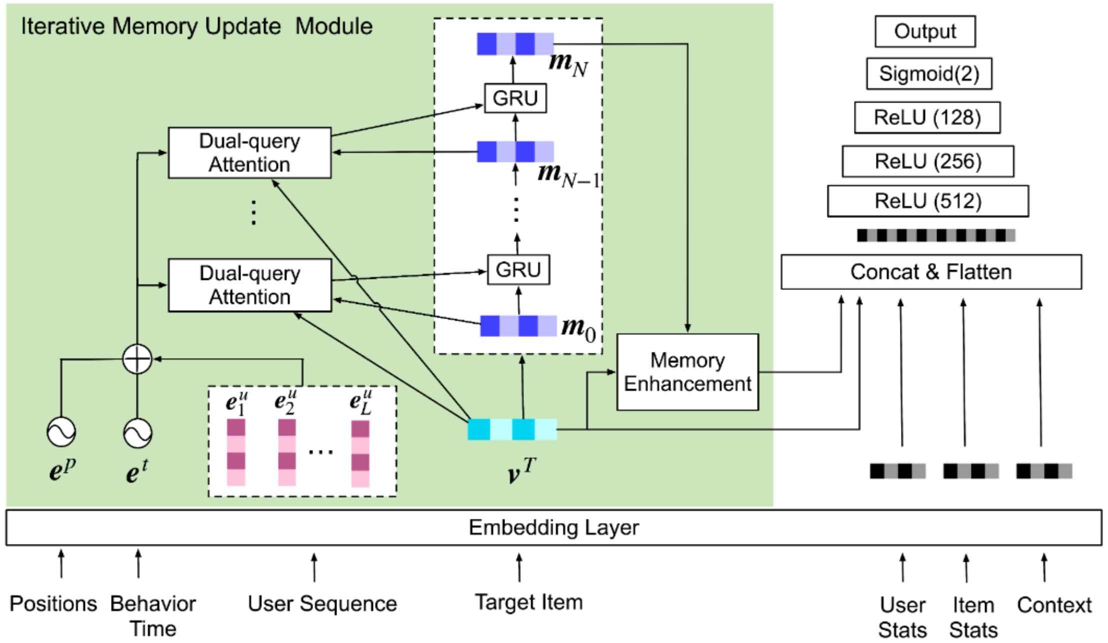
$$e(B^t) = \{e^t_1, e^t_2, ..., e^t_L\}$$

$$e(B^u) = \{e^u_1, e^u_2, ..., e^u_L\}$$

$$e^u_j = e^i_j \oplus e^t_j \oplus e^p_j$$

Method



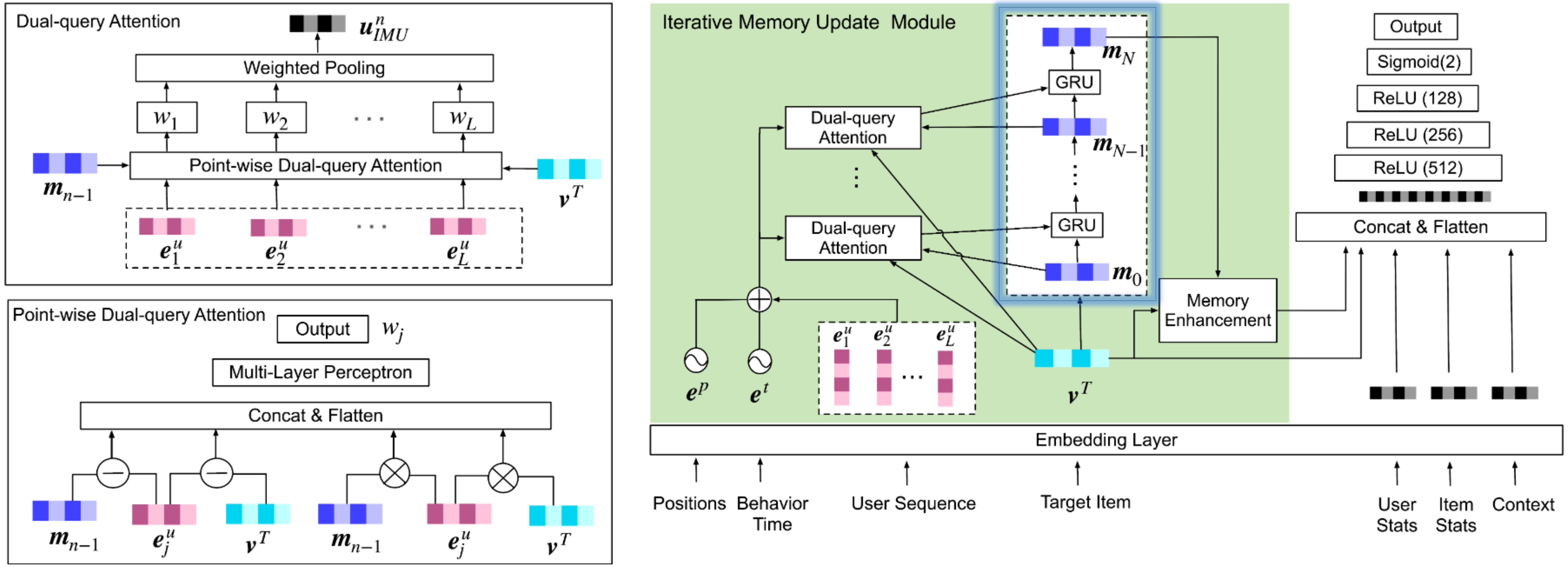


$$\alpha_j(e_j^u, v^T, m_t) = [e_j^u \ominus m_t \,\|\, e_j^u \ominus v^T \,\|\, e_j^u \otimes m_t \,\|\, e_j^u \otimes v^T] \quad (1)$$

$$a_j(e_j^u, v^T, m_t) = \sigma(W^{(2)}\sigma(W^{(1)}\alpha_j(e_j^u, v^T, m_t) + b^{(1)}) + b^{(2)}) \quad (2)$$

$$u_{IMU}^n = f(v^T, \cup(e(B^u)), m_{n-1}) = \sum_{j=1}^{L} a_j(e_j^u, v^T, m_{n-1}) e_j^u = \sum_{j=1}^{L} w_j e_j^u \quad (5)$$

Method

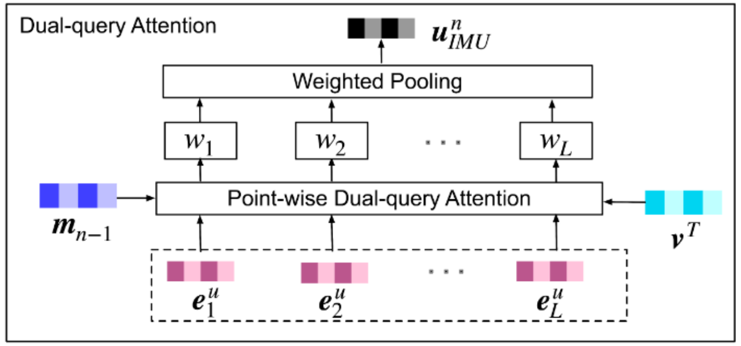


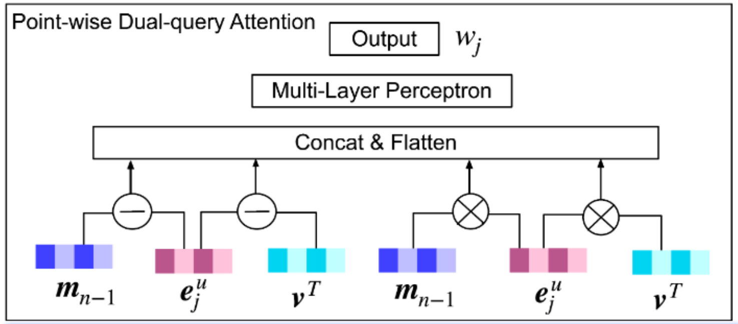
$$m_n \leftarrow f(\cup(e(B^u)), m_{n-1}) \quad (3)$$

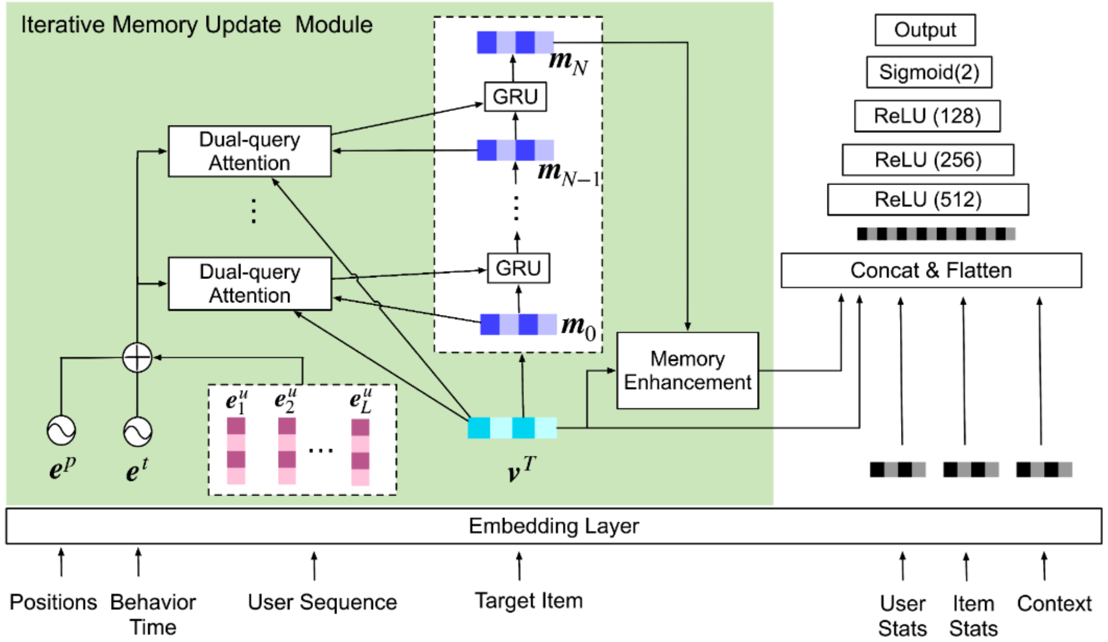
$$m_0 = v^T \quad (4)$$

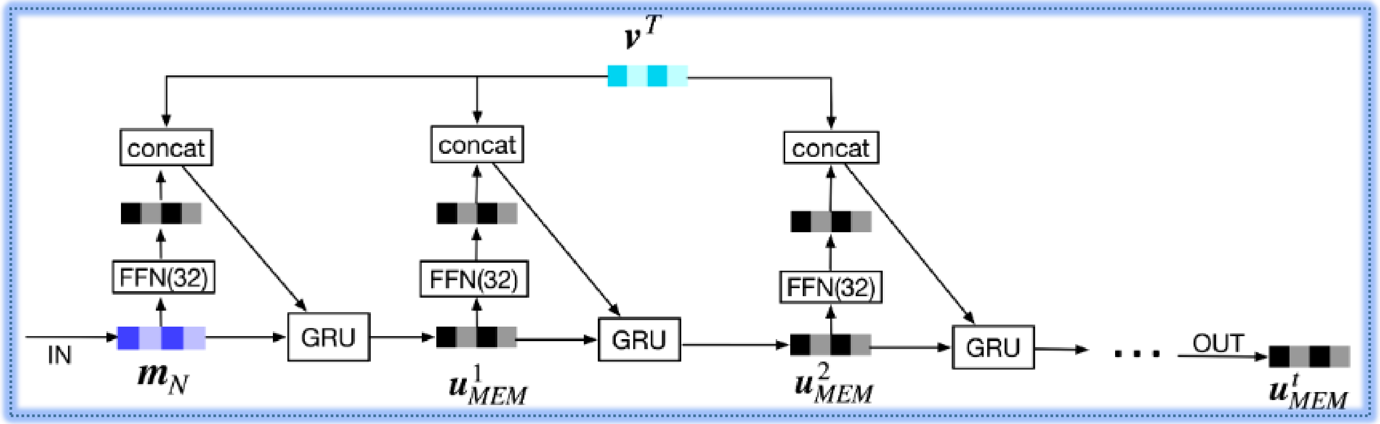
$$m_n = GRU(u^n_{IMU}, m_{n-1}) \quad (6)$$

Problem Statement









$$u^t_{MEM} = GRU([W^u u^{t-1}_{MEM} \,||\, v^T], u^{t-1}_{MEM}) \tag{7}$$

Experiments

	AUC (mean±std)		
	Books	**Movies**	**Industrial**
YouTube	0.83738(±0.00131)	0.83432(±0.00164)	0.73534(±0.000081)
DIN	0.85162(±0.00272)	0.86026(±0.00130)	0.73749(±0.000126)
DIEN	0.85498(±0.00128)	0.86542(±0.00072)	0.73807(±0.000093)
SASRec	0.82144(±0.00748)	0.83690(±0.00953)	0.73461(±0.000140)
MIMN	0.85228(±0.00138)	0.87140(±0.00085)	0.73678(±0.000201)
UBR4CTR	0.84834(±0.00062)	0.85957(±0.00145)	0.73649(±0.000096)
SAM 2P	0.85370(±0.00196)	0.88214(±0.00138)	0.73939(±0.000034)
SAM 3P	**0.86723(±0.00077)**	**0.88352(±0.00149)**	**0.74152(±0.000093)**
SAM 3P+	0.86926(±0.00142)	0.88628(±0.00097)	0.74234(±0.000087)
SAM 3P+ts	0.86997(±0.00113)	0.88714(±0.00157)	0.74238(±0.000103)

Table 1: Model performance (AUC) for two public benchmarks and the industrial dataset with maximum affordable sequence lengths.

Experiments

		YouTube	DIN	DIEN	SASRec	MIMN	UBR4CTR	**SAM 3P**
Books Dataset	SeqLen=50	0.80841	0.81873	0.84541	0.81008	0.82753	0.81762	**0.85662**
	SeqLen=100	0.81729	0.84569	0.84866	0.82144	0.84393	0.82833	**0.86056**
	SeqLen=200	0.82544	0.84724	0.85498	N.A.	0.85228	0.83488	**0.86377**
	SeqLen=500	0.83252	0.84807	N.A.	N.A.	N.A.	0.84165	**0.86538**
	SeqLen=1000	0.83738	0.85162	N.A.	N.A.	N.A.	0.84834	**0.86723**
Movies Dataset	SeqLen=50	0.81336	0.83538	0.84946	0.82978	0.85312	0.82824	**0.86347**
	SeqLen=100	0.82293	0.84676	0.85997	0.83690	0.86638	0.84297	**0.87032**
	SeqLen=200	0.82743	0.84917	0.86542	N.A.	0.87140	0.84739	**0.87691**
	SeqLen=500	0.83075	0.85563	N.A.	N.A.	N.A.	0.85301	**0.87950**
	SeqLen=1000	0.83432	0.86026	N.A.	N.A.	N.A.	0.85957	**0.88352**
Industrial Dataset	SeqLen=50	0.73019	0.73298	0.73304	0.73296	0.73212	0.73287	**0.73443**
	SeqLen=100	0.73236	0.73327	0.73331	0.73325	0.73379	0.73309	**0.73511**
	SeqLen=200	0.73264	0.73331	0.73599	0.73461	0.73678	0.73315	**0.73796**
	SeqLen=500	0.73371	0.73728	0.73807	N.A.	N.A.	0.73586	**0.74029**
	SeqLen=1000	0.73534	0.73749	N.A.	N.A.	N.A.	0.73649	**0.74152**

Table 2: Model performance (AUC) for varying sequence lengths for the proposed solution and the compared models. Experiments with N.A. incur Out-of-Memory (OOM) error during training.

Experiments

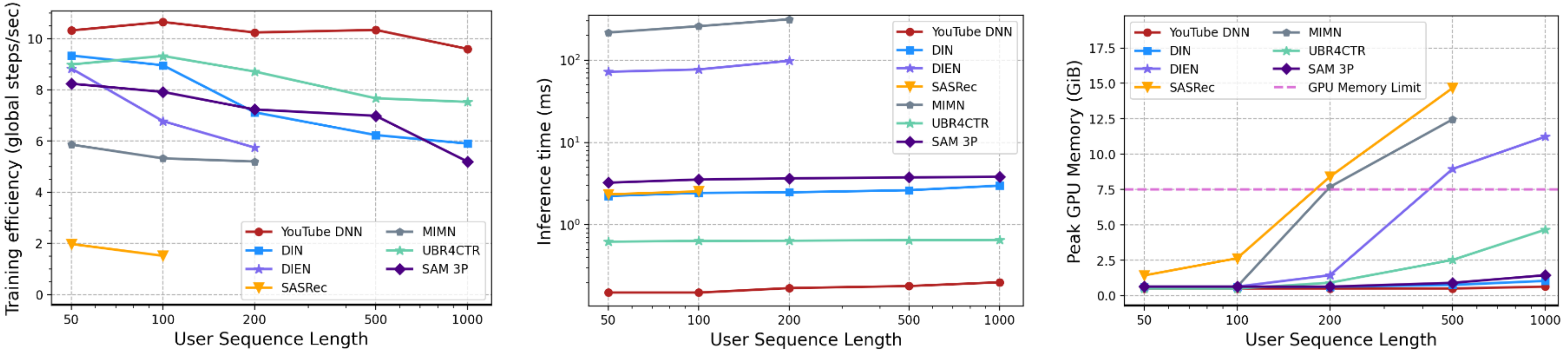
Method	AUC(mean±std)		
	Books	Movies	Industrial
w/o. attention	0.83738(±0.00131)	0.83432(±0.00164)	0.73534(±0.000081)
w/o. iterative walk	0.85162(±0.00272)	0.86026(±0.00130)	0.73749(±0.000126)
dot product	0.84885(±0.00147)	0.85644(±0.00115)	0.73677(±0.000085)
w/o. subtraction op.	0.86491(±0.00068)	0.87536(±0.00133)	0.74020(±0.000124)
delayed cross	0.85343(±0.00121)	0.86037(±0.00107)	0.73892(±0.000132)
w/o. m.e.	0.86723(±0.00077)	0.88352(±0.00149)	0.74152(±0.000093)
full (SAM 3P+ts)	0.86997(±0.00113)	0.88714(±0.00157)	0.74238(±0.000103)

Table 3: Ablation study on the SAM model structure

Method	Complexity	Seq. Op.	Max Path	Encoding
DIN	$O(L \cdot d)$	$O(1)$	$O(\infty)$	$(CROSS)$
DIEN	$O(L \cdot d^2)$	$O(L)$	$O(L)$	$(ENC, CROSS)$
SASRec	$O(L^2 \cdot d)$	$O(1)$	$O(1)$	$(ENC, CROSS)$
MIMN	$O(L \cdot d^2)$	$O(L)$	$O(L)$	$(ENC, CROSS)$
UBR4CTR	$O(L \cdot d)$	$O(1)$	$O(\infty)$	$(CROSS)$
SAM	$O(L \cdot d)$	$O(1)$	$O(1)$	$(CROSS, ENC)$

Table 4: Complexity, minimum number of sequential operations(abbreviated as Seq. Op.), maximum path length, and encoding paradigms for compared methods. L is the sequence length and d is the model dimension.

Experiments



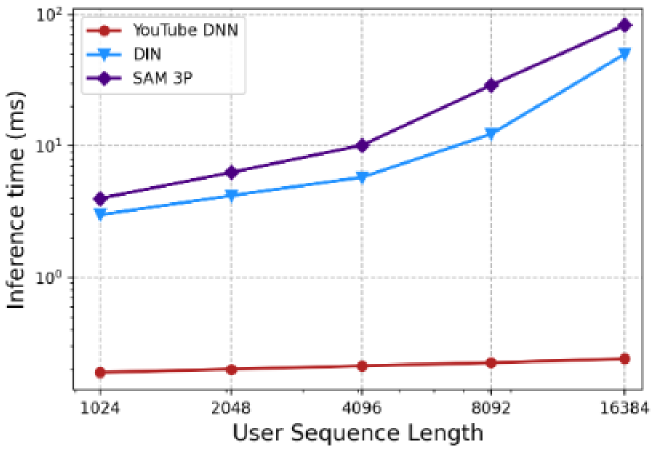
(a) Train Efficiency

(b) Inference Speed

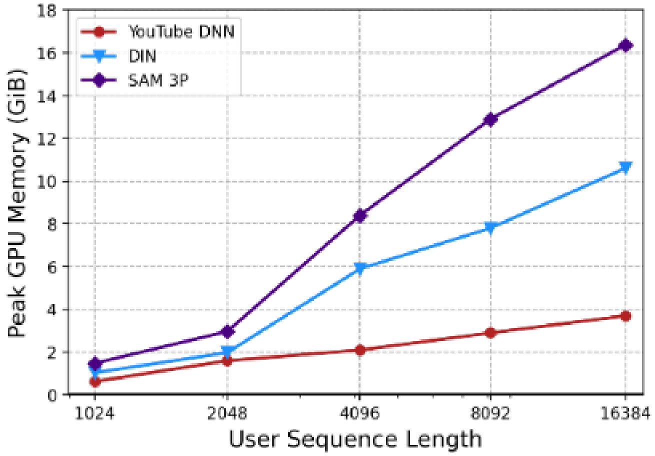
(c) Memory Efficiency

Figure 3: Computational cost and memory efficiency for all compared models. The x-axes are on logarithmic scales for all three plots. The y-axis for Fig.3b is on a logarithmic scale.

Experiments



(a) Inference Speed



(b) Memory Efficiency

Figure 4: Inference time and peak memory usage for extremely long sequences with lengths up to 16K. The y-axis for the inference time is on a logarithmic scale.

Experiments

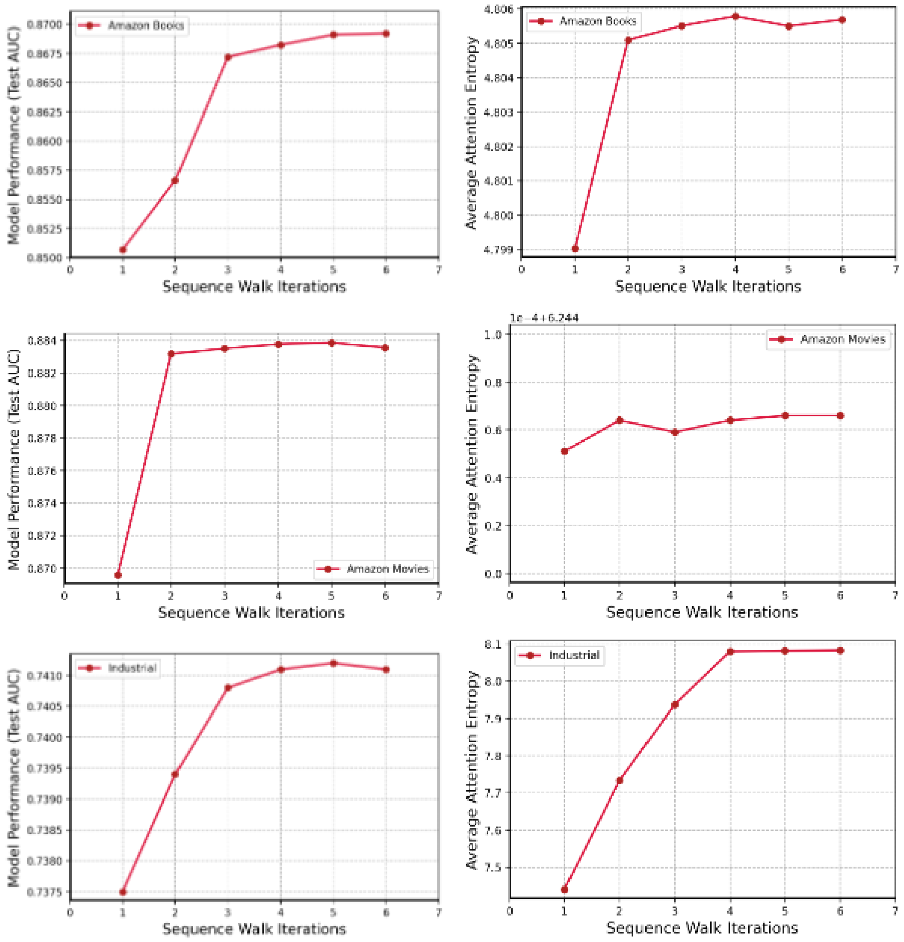


Figure 5: Model performance (AUC) and the entropy of the attention distribution against memory update iterations.

$$Entropy_{\alpha}(x) = -\sum_{i}^{L}(\alpha_i(x)\log(\alpha_i(x))) \qquad (8)$$

Table 5: Online A/B test results for consecutive 9 days. The row Impr denotes relative improvement.

	Online A/B Metrics (mean±std)		
	CTR	TCIC	CCC
Base	4.4254%±0.0244%	325741±2701.9	2.979±0.0133
SAM	4.7482%±0.0222%	375733±4055.1	3.193±0.0181
Impr	**7.30%±0.93%**	**15.36%±1.65%**	**7.19%±0.80%**

Thanks