SDM: Sequential Deep Matching Model for Online Large-scale Recommender System

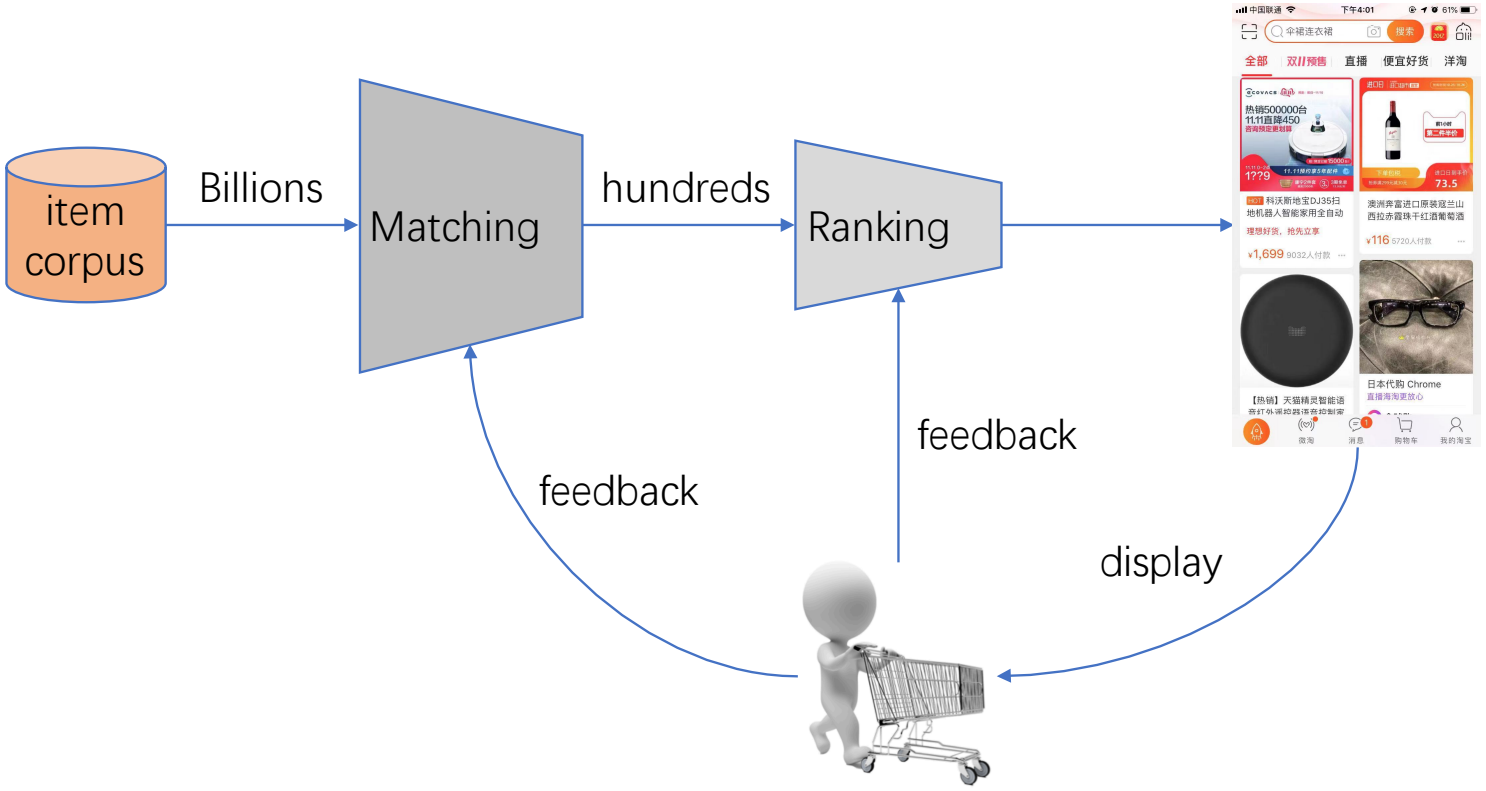
Fuyu Lv[1], Taiwei Jin[1], Changlong Yu[2], Fei Sun[1], Quan Lin[1], Keping Yang[1], Wilfred Ng[2]

[1]Search & Recommendation, Alibaba Group

[2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology

Background

- Two stages in recommender system at Taobao: **Matching** and **Ranking**
- Matching: retrieve a candidate set of items for ranking



Background

- Online deployed matching: item similarity based CF

- Can not well capture dynamic transformation of user interests

Background

- We introduce deep sequential recommendation model in matching stage
- Matching is measured by inner products of user behavior and item vectors
- Sequential user behaviors are modeled from long-term and short-term views
  - Short-term sessions represent current shopping demands
  - Long-term behaviors represent historical general preferences

Problem Formulation

- Given short-term sessions $\mathcal{S}^u$ and long-term behaviors $\mathcal{L}^u$
- Predict next Top N interacted items

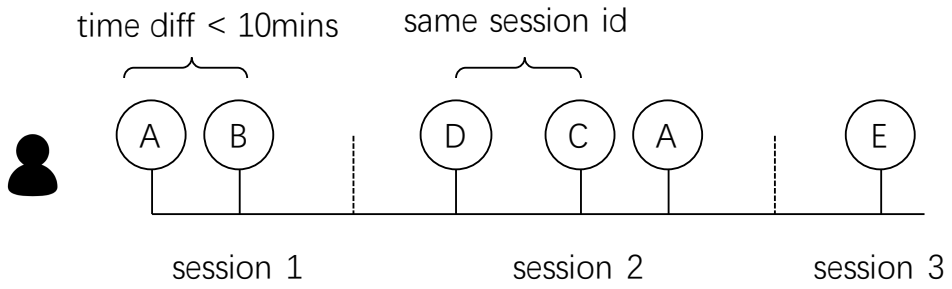
$$P(i|\mathcal{S}^u, \mathcal{L}^u; \theta)$$

- Tackle two inherent problems
  - Multiple user interests in one short-term session

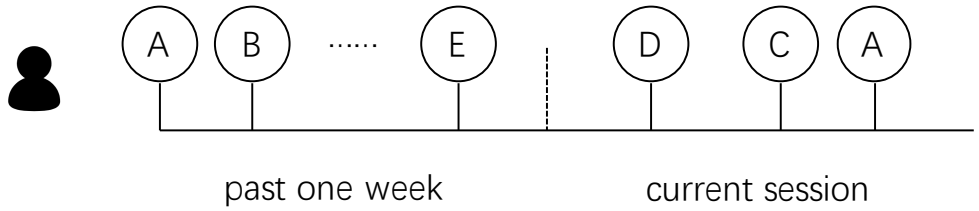
  - Effectively combine short-term and long-term representations

Sequence Generation

- Short-term sessions
  - session id + time

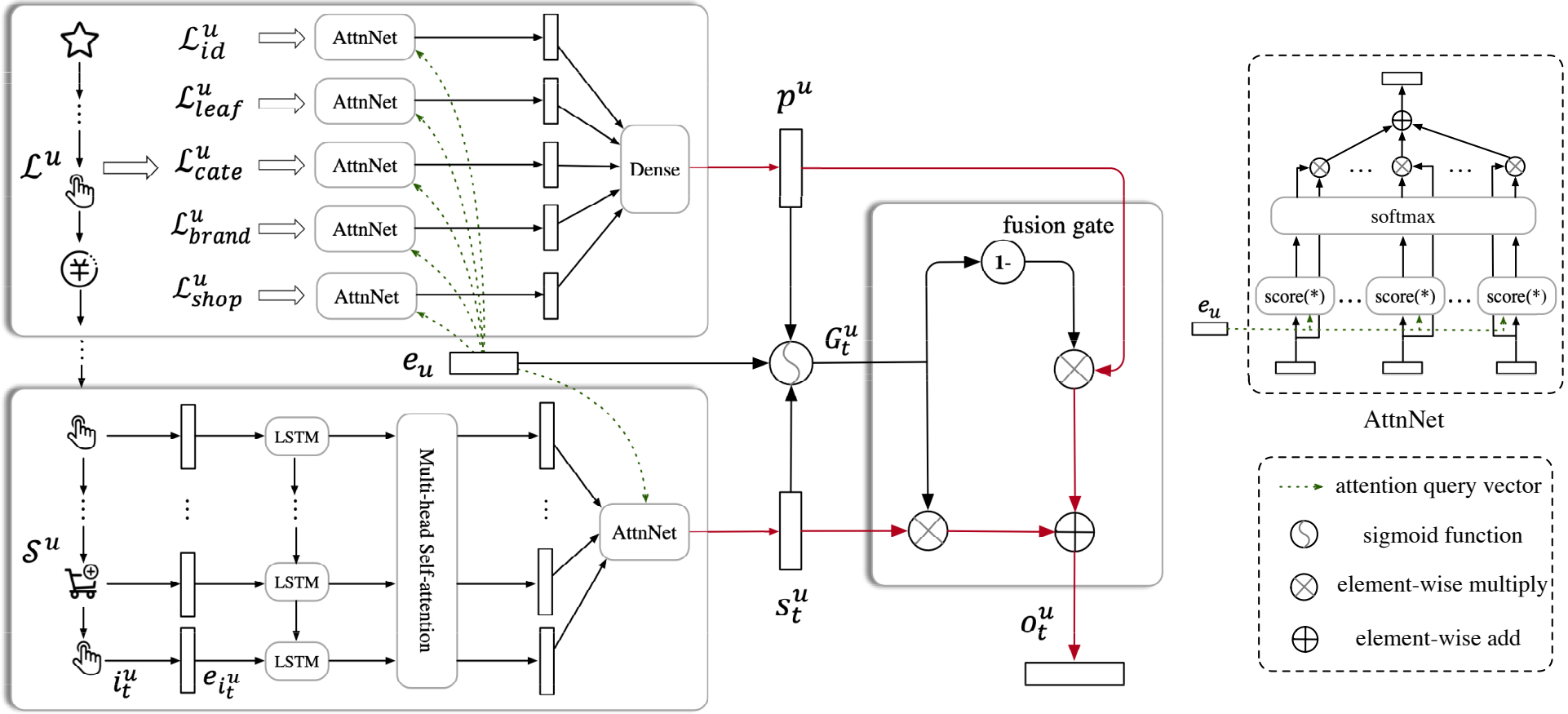


- Long-term behaviors
  - past one week



Model Architecture

- Overview



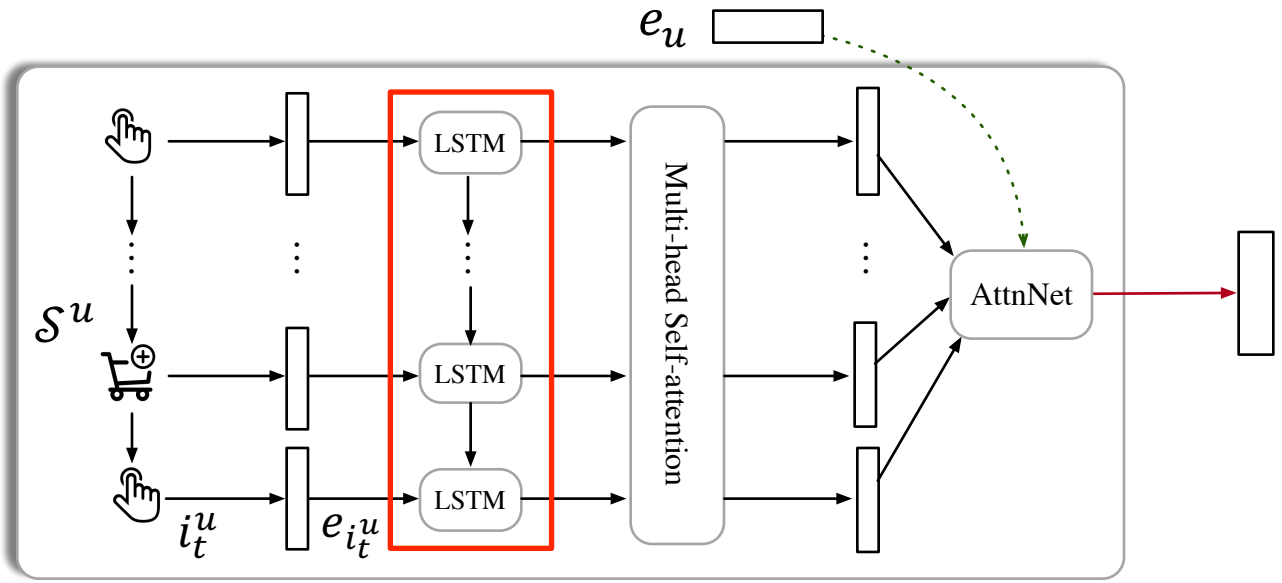
Model Architecture

- Short-term session modeling
  - Item embedding with side information

set of side information

$$\boldsymbol{e}_{i_t^u} = \text{concat}(\{\boldsymbol{e}_i^f \mid f \in \mathcal{F}\})$$

  - LSTM: dynamic user interests

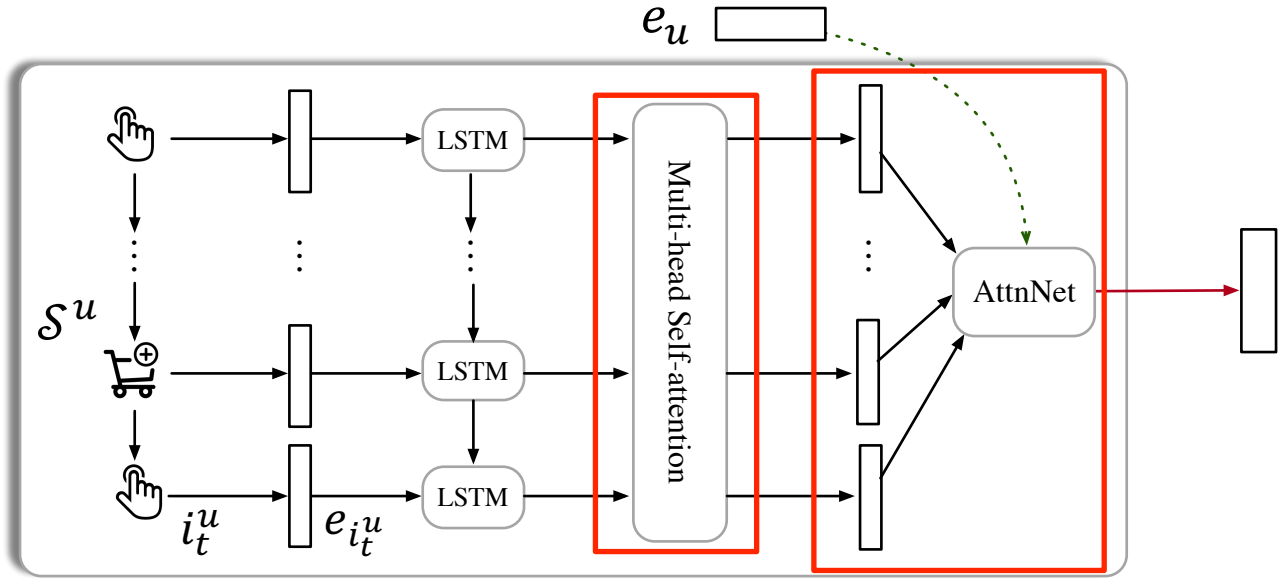


Model Architecture

- Short-term session modeling
  - Multi-head Self-attention: multiple user interests
    - single interest i, at time t $\quad \mathbf{head}_i^u = \text{Attention}(\boldsymbol{W}_i^Q \boldsymbol{X}^u, \boldsymbol{W}_i^K \boldsymbol{X}^u, \boldsymbol{W}_i^V \boldsymbol{X}^u)$
    - overall interest $\quad \hat{\boldsymbol{X}}^u = \text{MultiHead}(\boldsymbol{X}^u) = \boldsymbol{W}^O \text{concat}(\mathbf{head}_1^u, \ldots, \mathbf{head}_h^u)$

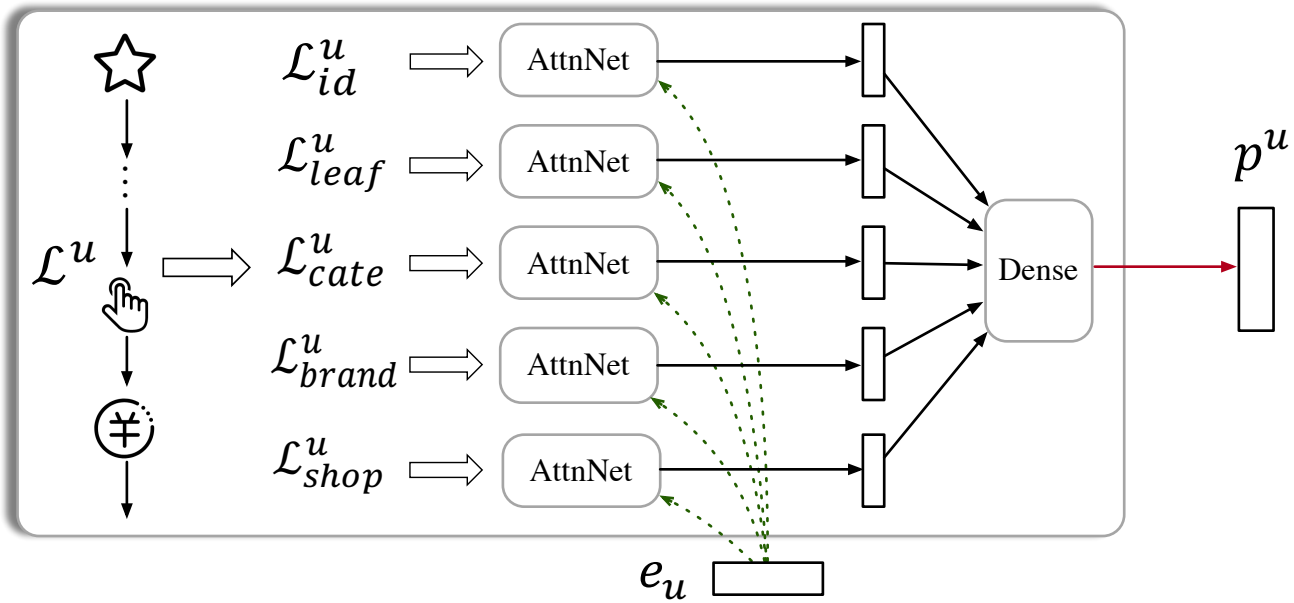
Outputs of LSTM

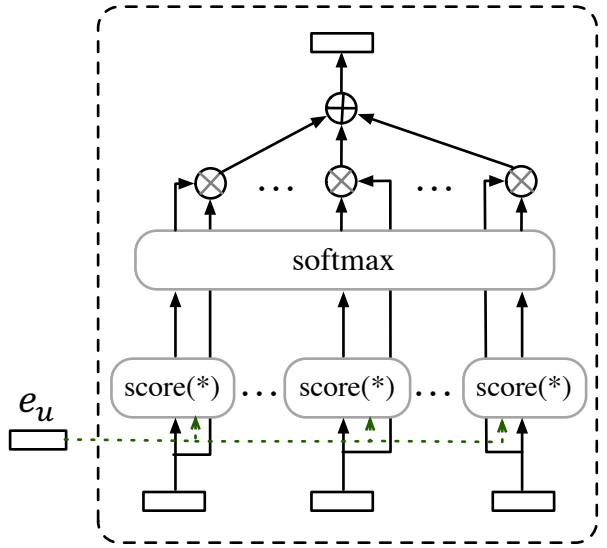
  - User attention
    - personalized
    - aggregate interests over time $1 \ldots t$



Model Architecture

- Long-term behaviors modeling
  - General interests of different feature scales





AttnNet

$$z^u = \text{concat}(\{z_f^u | f \in \mathcal{F}\})$$

$$p^u = \tanh(W^p z^u + b)$$

$$\alpha_k = \frac{\exp(g_k^{uT} e_u)}{\sum_{k=1}^{|\mathcal{L}_f^u|} \exp(g_k^{uT} e_u)}$$

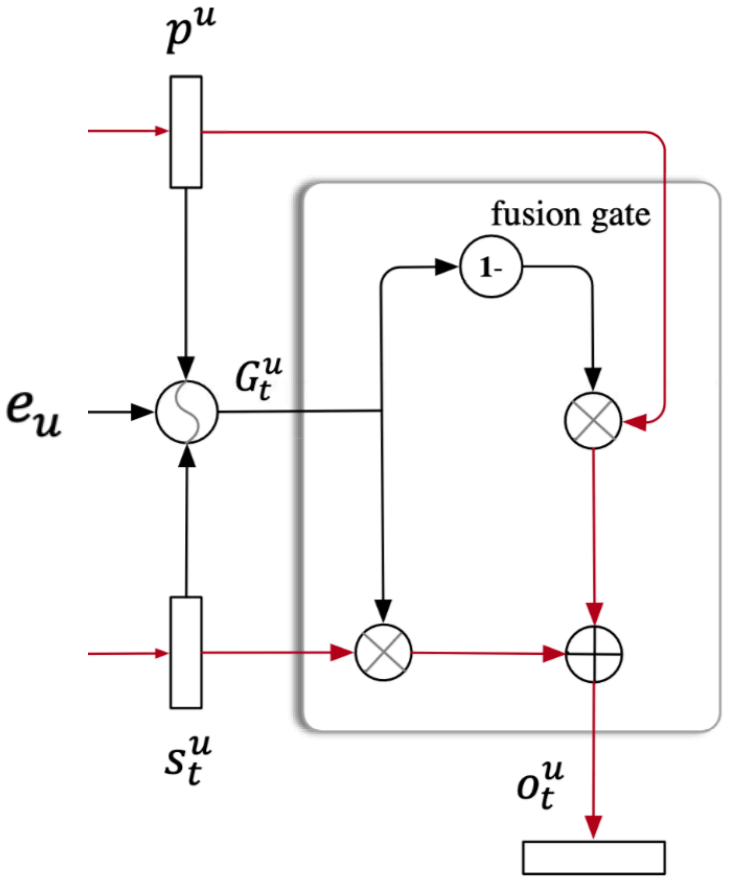
$$z_f^u = \sum_{k=1}^{|\mathcal{L}_f^u|} \alpha_k g_k^u$$

Model Architecture

- Long-term behaviors fusion
  - Gate Mechanism

$$G_t^u = \text{sigmoid}(W^1 e_u + W^2 s_t^u + W^3 p^u + b)$$

$$o_t^u = (1 - G_t^u) \odot p^u + G_t^u \odot s_t^u$$



Model Architecture

- Training
  - the next interacted item as positive label
  - sample K items from item corpus as negative labels
  - scores between user behaviors and item vectors

item vectors

$$z_i = \text{score}(\boldsymbol{o}_t^u, \boldsymbol{v_i}) = \boldsymbol{o}_t^{uT} \boldsymbol{v_i}$$

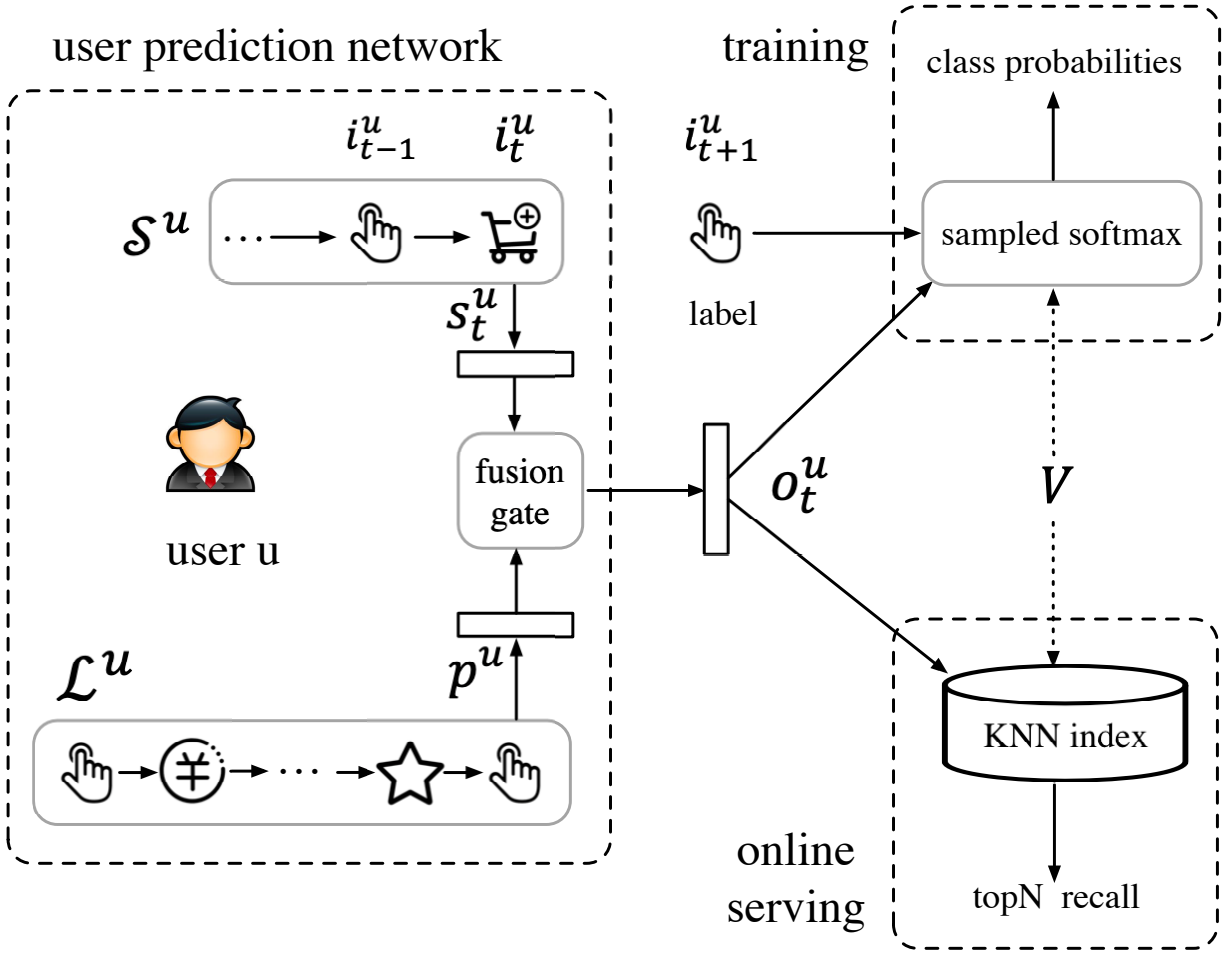
  - softmax and minimize cross-entropy

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{z})$$

$$L(\hat{\boldsymbol{y}}) = -\sum_{i \in \mathcal{K}} y_i \log(\hat{y}_i)$$

System Deployment

- Framework



Experiments

- Datasets
  - E-commerce

Dataset	Data Type	Data Split	#[a]Users	#Items	#Records	#Sessions	S.len[b]	L.size[c]	Time Interval
JD	offline	train	802,479	154,568	9,653,777	2,666,189	3.3	20	15/Mar/2018 - 8/Apr/2018
		test	10,366	74,564	498,492	15,069	8.6	20	9/Apr/2018 - 15/Apr/2018
Taobao	offline	train	498,633	2,053,798	45,157,298	7,011,385	6.1	20	15/Dec/2018 - 21/Dec/2018
		test	13,237	588,306	1,170,401	13,237	9.2	20	22/Dec/2018
	online	train	3.3×10^8	1×10^8	2.1×10^{10}	2.7×10^9	7.1	50	Dec/2018
		test	3.3×10^8	1×10^8	/	/	8.1	50	Dec/2018

[a] # means the number of. [b] S.len is the average length of short-term behaviors. [c] L.size is the maximum size of each subset in long-term behaviors.

- Evaluations
  - Offline

$$\text{HitRate@K} = \frac{n_{hit}}{N} \qquad \text{Precision@K}(u) = \frac{|P_u \cap G_u|}{K} \qquad \text{Recall@K}(u) = \frac{|P_u \cap G_u|}{|G_u|}$$

  - Online

$$\text{pCTR} = \frac{\#\text{clicks}}{\#\text{pages}} \qquad \text{pGMV} = 1000 \times \frac{\#\text{pay amount}}{\#\text{pages}} \qquad \text{discovery} = \frac{\#\text{new categories}}{\#\text{all categories}}$$

Experiments

- Model Comparisons
  - Overall Performance

Models	Taobao				JD			
	HitRate@100	Recall@100	Precision@100	F1@100	HitRate@20	Recall@20	Precision@20	F1@20
Item-based CF	60.27%	3.24%	2.00%	2.43%	67.50%	9.08%	9.41%	8.99%
DNN	60.88%	2.85%	1.83%	2.18%	68.43%	8.93%	9.65%	8.98%
GRU4REC	65.60%	3.66%	2.30%	2.77%	69.44%	9.33%	9.83%	9.29%
NARM	66.97%	3.57%	2.25%	2.70%	70.33%	9.07%	9.58%	9.04%
SHAN	67.30%	3.71%	2.33%	2.80%	70.54%	9.42%	10.02%	9.41%
BINN	67.55%	3.49%	2.20%	2.64%	72.19%	9.38%	9.93%	9.36%
SDMMA	68.24%	3.68%	2.32%	2.79%	70.41%	9.21%	9.72%	9.18%
PSDMMA	69.43%	3.75%	2.37%	2.84%	71.21%	9.21%	9.78%	9.20%
PSDMMAL	70.72%	**3.86%**	2.44%	**2.93%**	73.25%	9.47%	10.13%	9.48%
PSDMMAL-N	**73.13%**	3.83%	**2.45%**	2.92%	**74.33%**	**9.68%**	**10.42%**	**9.72%**
PSDMMAL-NoS	65.41%	3.38%	2.14%	2.56%	70.07%	9.05%	9.60%	9.03%

Sequential recommendation models can beat item similarity based CF and DNN

Experiments

- Model Comparisons
  - Overall Performance

Models	Taobao				JD			
	HitRate@100	Recall@100	Precision@100	F1@100	HitRate@20	Recall@20	Precision@20	F1@20
Item-based CF	60.27%	3.24%	2.00%	2.43%	67.50%	9.08%	9.41%	8.99%
DNN	60.88%	2.85%	1.83%	2.18%	68.43%	8.93%	9.65%	8.98%
GRU4REC	65.60%	3.66%	2.30%	2.77%	69.44%	9.33%	9.83%	9.29%
NARM	66.97%	3.57%	2.25%	2.70%	70.33%	9.07%	9.58%	9.04%
SHAN	67.30%	3.71%	2.33%	2.80%	70.54%	9.42%	10.02%	9.41%
BINN	67.55%	3.49%	2.20%	2.64%	72.19%	9.38%	9.93%	9.36%
SDMMA	68.24%	3.68%	2.32%	2.79%	70.41%	9.21%	9.72%	9.18%
PSDMMA	69.43%	3.75%	2.37%	2.84%	71.21%	9.21%	9.78%	9.20%
PSDMMAL	70.72%	**3.86%**	2.44%	**2.93%**	73.25%	9.47%	10.13%	9.48%
PSDMMAL-N	**73.13%**	3.83%	**2.45%**	2.92%	**74.33%**	**9.68%**	**10.42%**	**9.72%**
PSDMMAL-NoS	65.41%	3.38%	2.14%	2.56%	70.07%	9.05%	9.60%	9.03%

Enhanced by multi-head attention, SDM performs the best than GRU-based sequential models in terms of short-term session modeling

Experiments

- Model Comparisons
  - Overall Performance

Models	Taobao				JD			
	HitRate@100	Recall@100	Precision@100	F1@100	HitRate@20	Recall@20	Precision@20	F1@20
Item-based CF	60.27%	3.24%	2.00%	2.43%	67.50%	9.08%	9.41%	8.99%
DNN	60.88%	2.85%	1.83%	2.18%	68.43%	8.93%	9.65%	8.98%
GRU4REC	65.60%	3.66%	2.30%	2.77%	69.44%	9.33%	9.83%	9.29%
NARM	66.97%	3.57%	2.25%	2.70%	70.33%	9.07%	9.58%	9.04%
SHAN	67.30%	3.71%	2.33%	2.80%	70.54%	9.42%	10.02%	9.41%
BINN	67.55%	3.49%	2.20%	2.64%	72.19%	9.38%	9.93%	9.36%
SDMMA	68.24%	3.68%	2.32%	2.79%	70.41%	9.21%	9.72%	9.18%
PSDMMA	69.43%	3.75%	2.37%	2.84%	71.21%	9.21%	9.78%	9.20%
PSDMMAL	70.72%	**3.86%**	2.44%	**2.93%**	73.25%	9.47%	10.13%	9.48%
PSDMMAL-N	**73.13%**	3.83%	**2.45%**	2.92%	**74.33%**	**9.68%**	**10.42%**	**9.72%**
PSDMMAL-NoS	65.41%	3.38%	2.14%	2.56%	70.07%	9.05%	9.60%	9.03%

1. Our model can beat SHAN and BINN, which also use long-term behaviors
2. Multi-head attention and gate mechanism ensure the superior performance

Experiments

- Model Comparisons
  - Overall Performance

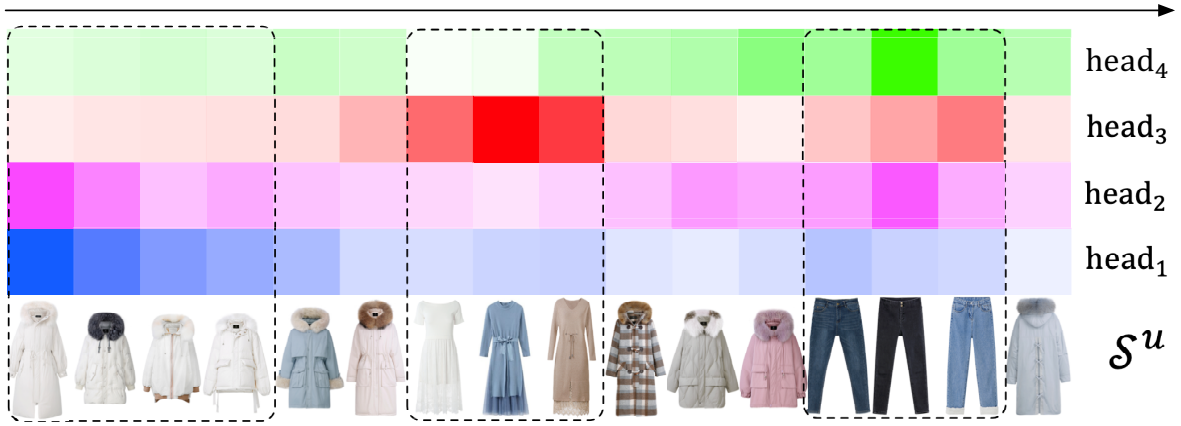
Models	Taobao				JD			
	HitRate@100	Recall@100	Precision@100	F1@100	HitRate@20	Recall@20	Precision@20	F1@20
Item-based CF	60.27%	3.24%	2.00%	2.43%	67.50%	9.08%	9.41%	8.99%
DNN	60.88%	2.85%	1.83%	2.18%	68.43%	8.93%	9.65%	8.98%
GRU4REC	65.60%	3.66%	2.30%	2.77%	69.44%	9.33%	9.83%	9.29%
NARM	66.97%	3.57%	2.25%	2.70%	70.33%	9.07%	9.58%	9.04%
SHAN	67.30%	3.71%	2.33%	2.80%	70.54%	9.42%	10.02%	9.41%
BINN	67.55%	3.49%	2.20%	2.64%	72.19%	9.38%	9.93%	9.36%
SDMMA	68.24%	3.68%	2.32%	2.79%	70.41%	9.21%	9.72%	9.18%
PSDMMA	69.43%	3.75%	2.37%	2.84%	71.21%	9.21%	9.78%	9.20%
PSDMMAL	70.72%	**3.86%**	2.44%	**2.93%**	73.25%	9.47%	10.13%	9.48%
PSDMMAL-N	**73.13%**	3.83%	**2.45%**	2.92%	**74.33%**	**9.68%**	**10.42%**	**9.72%**
PSDMMAL-NoS	65.41%	3.38%	2.14%	2.56%	70.07%	9.05%	9.60%	9.03%

When N = 5 and combining short and long-term behaviors, our model can achieve the best performance

Experiments

- Model Analysis
  - Effectiveness of Multi-head

#heads	HitRate@K	Recall@K	Precision@K	F1@k
1	70.00%	3.82%	2.40%	2.88%
2	70.64%	3.83%	2.41%	2.89%
4	**70.72%**	**3.86%**	**2.44%**	**2.93%**
8	70.21%	3.77%	2.37%	2.85%



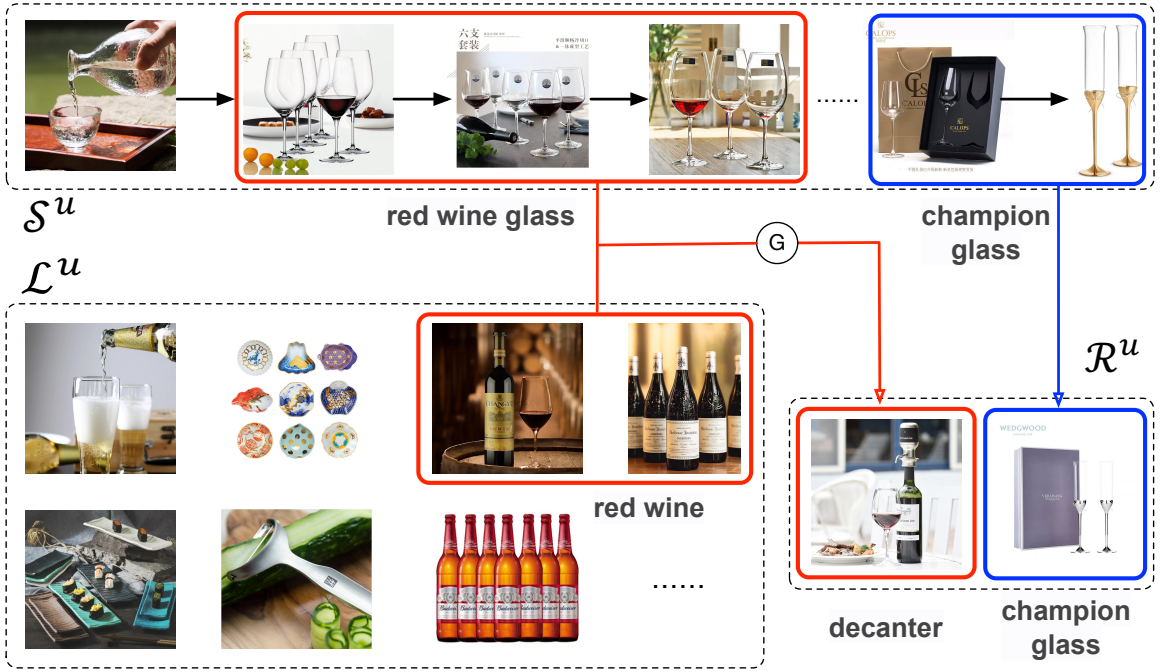
Different session interests captured by different heads

1. Influence of multi-head as the number of heads increasing
2. Four heads gets the best results when d = 64

Experiments

- Model Analysis
  - Effectiveness of Gate

fusion	HitRate@K	Recall@K	Precision@K	F1@K
multiply	67.09%	3.42%	2.16%	2.59%
concat	69.74%	3.70%	2.34%	2.80%
add	70.24%	3.75%	2.37%	2.84%
gated	**70.72%**	**3.86%**	**2.44%**	**2.93%**

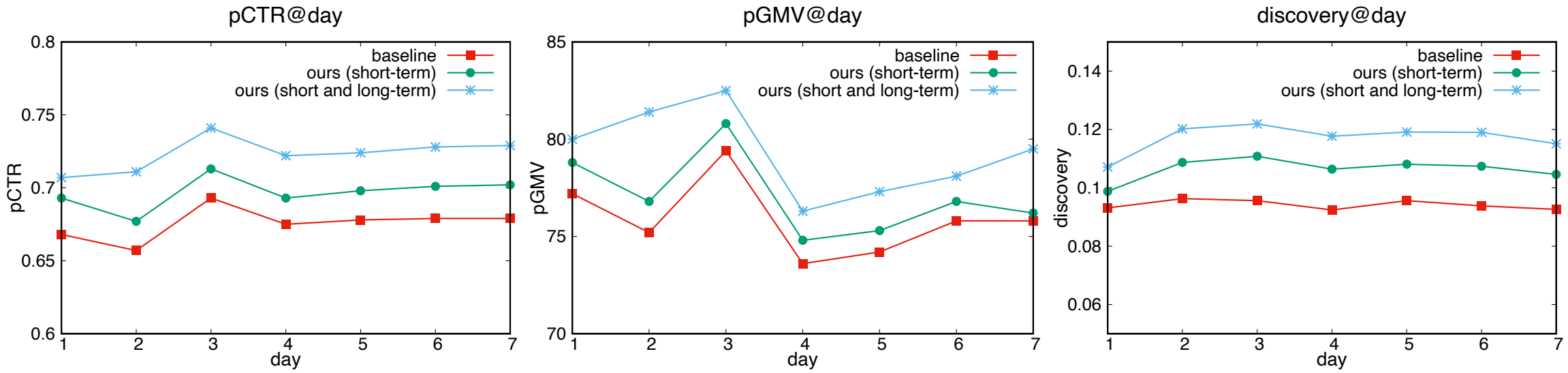


Gate fusion outperforms simple combination methods

1. The gate mechanism allows the information of present shopping demands remain, and
2. combine with long-term preferences to generate novel items for recommendations.

Experiments

- Online A/B Testing



- Fully deployed since December 2018
  - merged with item similarity based CF, +10% CTR

Conclusion

- We **develop**
  - a novel sequential deep matching (SDM) model for large-scale recommender system in real-world applications
- We **propose**
  - to model short-term sessions by multi- head self-attention to capture multiple interest tendencies, and a gated fusion to effectively combine long-term preferences
- We **deploy**
  - SDM on production environment of recommender system at Taobao with significant improvements of commercial metrics

Thank you
&
Question