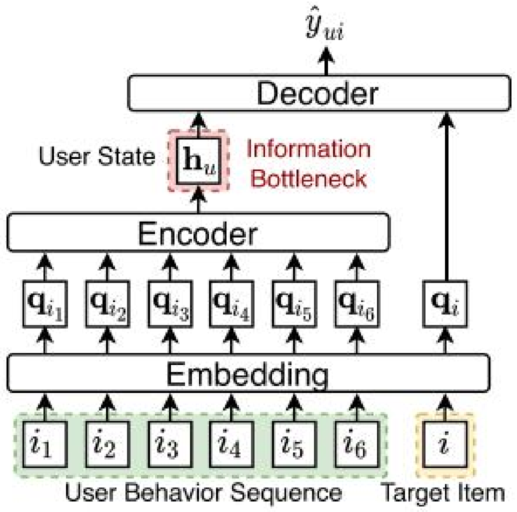
Seq2Bubbles: Region-Based Embedding Learning for User Behaviors in Sequential Recommenders

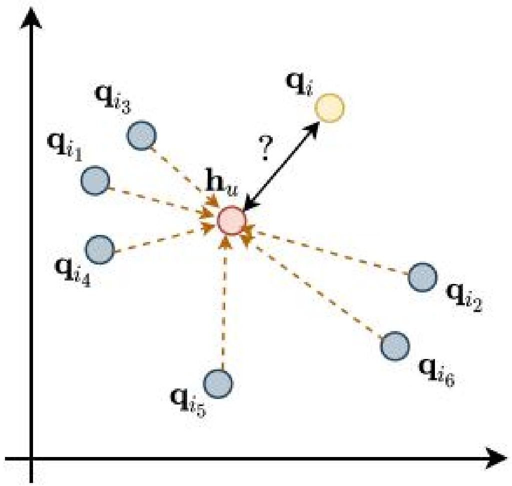
Qitian Wu, Chenxiao Yang, Shuodian Yu, Xiaofeng Gao, Guihai Chen
Shanghai Jiao Tong University

Background for Recommendation

- ❑ **Predict the next item based on historically clicked items of the user**
- ❑ **Most existing sequential recommendation models:**
  - *I. Embedding:* transform the item sequence into a sequence of vectors
  - *II. Encoding:* encode the sequence to estimate user interests
  - *III. Decoding:* compute similarity between the user state and a target item



(a)



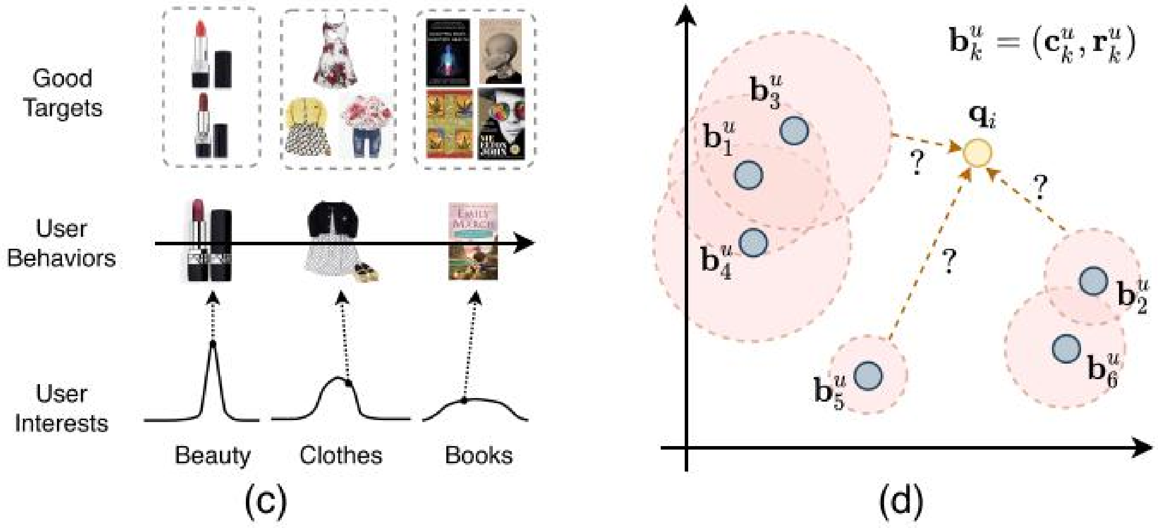
(b)

Squash a high-dimensional sequence into a single point

Motivation

- ❑ User interests often distribute over items of different aspects
  - Distribution of user interest tends to be multi-modal
- ❑ User interests for different items have distinct concentration levels
  - user's concentration: variance of user's clicked items in a specific aspect
  - more (less) diverse items in the aspect with stronger (weaker) concentration

Traditional point embedding fails to capture such distinct concentration levels!



Our Solutions: Region-based Embedding

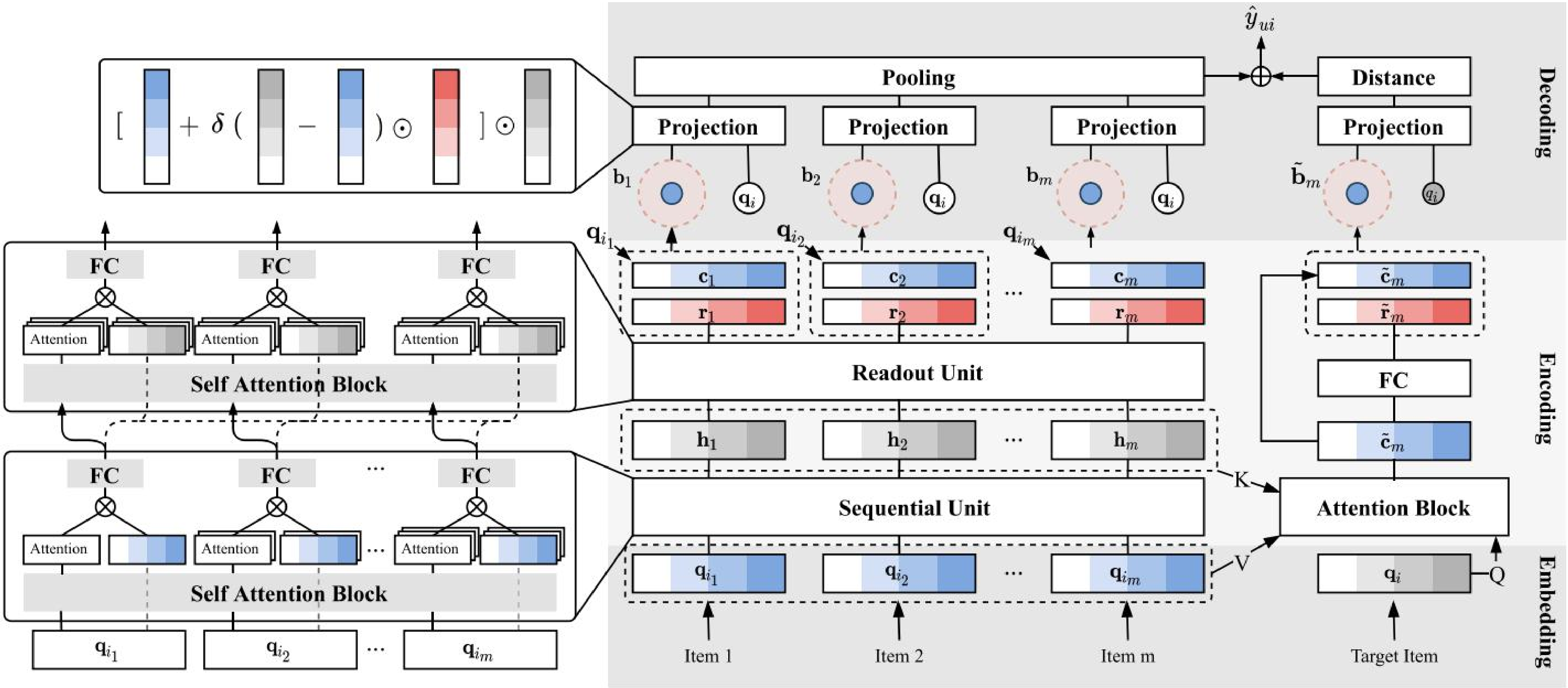
- Basic idea: embed a sequence into a set of bubbles
  - a hyper-ellipsoid in vector space
  - bubble center: clicked item embedding
  - bubble radius: embody concentration of user interests
  - a union of bubble embedding for sequence reflect user interests

$$\bigcup_{k=1}^{m} \{\mathbf{x} : \|(\mathbf{x} - \mathbf{c}_k) \odot \frac{1}{\mathbf{r}_k}\|_2 \leq 1\}$$

- Advantages:
  - Superior Expressiveness
  - Enough Flexibility
  - Interpretability

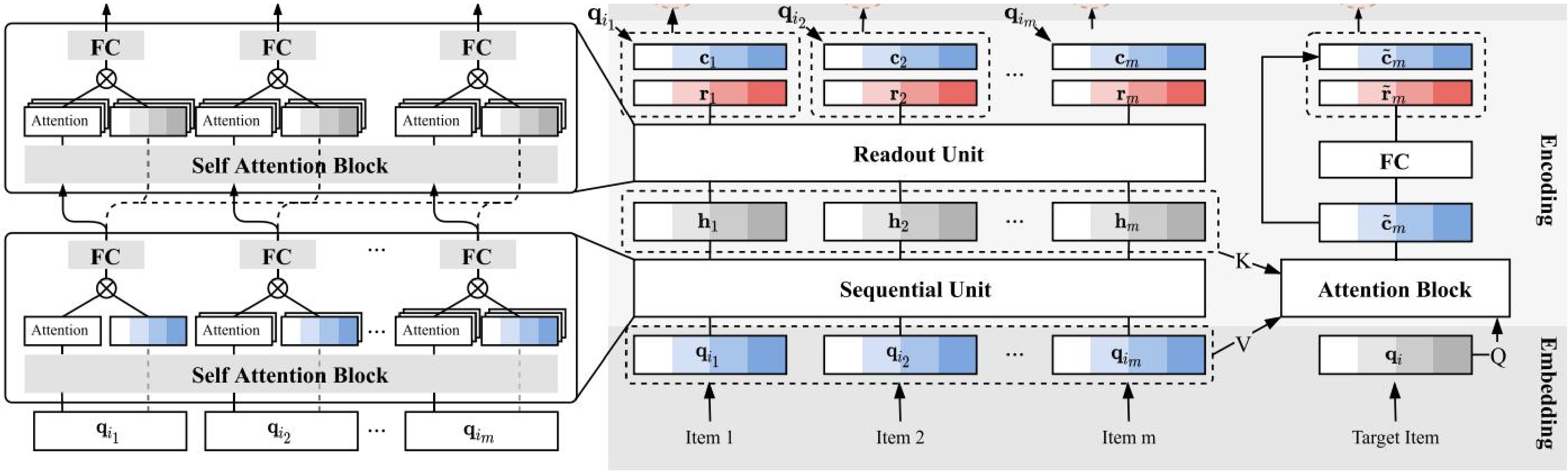
Key insight: regions enclosed by bubbles represent multi-modal interest and user intent

Proposed Model Overview



Model: Encoding Layer

- ❑ Encode item embedding sequence to extract useful information:
  - Filter out noise existing in behavior sequences
  - Mine temporal dependency and user's interests evolution
  - Distinguish the importance of different historical behaviors



Model: Encoding Layer (cont.)

❑ Self-attentive architecture:

- Lower-level sequential unit $\Phi_A(\cdot)$ to aggregate historical items

$$\mathbf{z}_k = \sum_{j=1}^{k} \alpha_{jk} \mathbf{q}_{i_j}, \quad \text{where } \alpha_{jk} = \sigma\left(\frac{(\mathbf{W}_K^1 \mathbf{q}_{i_k})^\top (\mathbf{W}_Q^1 \mathbf{q}_{i_j})}{\sqrt{d}}\right) \qquad \mathbf{h}_k = Dropout(PReLU(\mathbf{W}_N^1 \mathbf{z}_k + \mathbf{b}_N^1))$$

- High-level readout unit $\Phi_R(\cdot)$ to estimate radius of bubbles

$$\mathbf{z}_k = \sum_{j=1}^{m} \beta_{jk} \cdot \mathbf{h}_j, \quad \text{where } \beta_{jk} = \sigma\left(\frac{(\mathbf{W}_K^2 \mathbf{h}_k)^\top (\mathbf{W}_Q^2 \mathbf{h}_j)}{\sqrt{d}}\right) \qquad \mathbf{r}_k = Softplus(\mathbf{W}_N^2 \mathbf{z}_k + \mathbf{b}_N^2), \quad k = 1, \cdots, m$$

Model: Decoding Layer

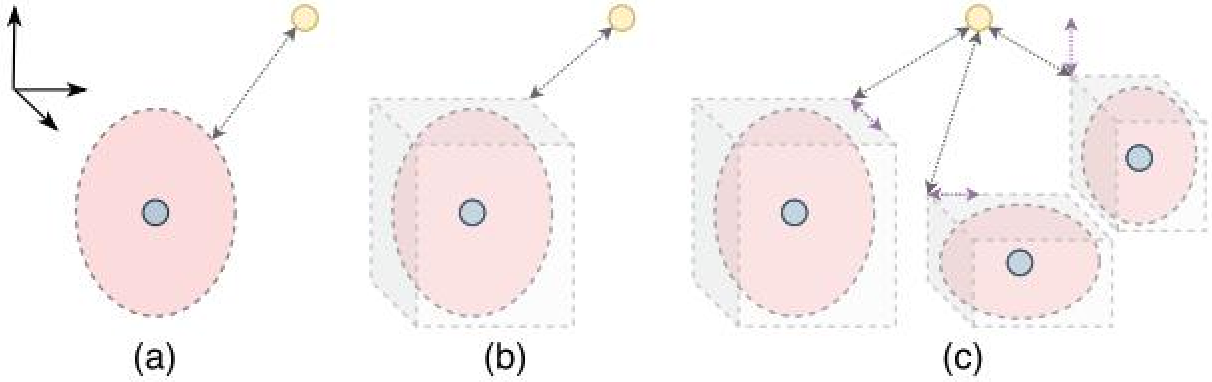
- **Compute the similarity between bubble embedding and target item**
  - ↪ the distance from a point to the surface of a hyper-ellipsoid?
- **Approximation:**
  - Consider a circumscribed hyper-cube outside the hyper-ellipsoid region

$$\mathbf{b} = \{\mathbf{c}, \mathbf{r}\}: [c_1 - r_1, c_1 + r_1] \times \cdots \times [c_d - r_d, c_d + r_d]$$

$$D(\mathbf{b}, \mathbf{q}) := \min_{\mathbf{e} \in \{-1,1\}^d} d(\mathbf{c} + \mathbf{e} \odot \mathbf{r}, \mathbf{q})$$

$$\mathcal{D}(\mathcal{B}^m, \mathbf{q}_i) := \min_{1 \le k \le m} D(\mathbf{b}_k, \mathbf{q}_i),$$
$$= \min_{1 \le k \le m} d(\mathbf{c}_k + \delta(\mathbf{q}_i - \mathbf{c}_k) \odot \mathbf{r}_k, \mathbf{q}_i)$$

$$\mathcal{S}(\mathcal{B}^m, \mathbf{q}_i) = \max_{1 \le k \le m} s(\mathbf{c}_k + \delta(\mathbf{q}_i - \mathbf{c}_k) \odot \mathbf{r}_k, \mathbf{q}_i)$$

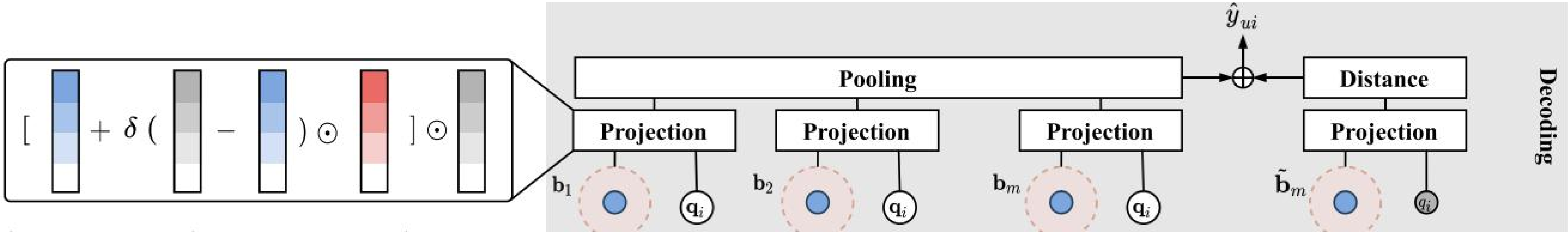


Model: Decoding Layer (cont.)

- ❑ Maximum operation only selects one bubble
  - The gradient only update one item
  - Ignore effects from different feature dimensions
- ❑ A generalized version:
  - max-pooling to select dominant bubbles in each feature dimension

$$\mathbf{p}_k = [\mathbf{c}_k + \delta(\mathbf{q}_i - \mathbf{c}_k) \odot \mathbf{r}_k] \odot \mathbf{q}_i, \quad k = 1, \cdots, m,$$
$$\mathbf{a}_m = \mathrm{MaxPooling}\{[\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_m]\}$$
$$\mathcal{S}(\mathcal{B}_m, \mathbf{q}_i) = s(\mathbf{a}_m, \mathbf{q}_i).$$

$$\hat{y}_{ui}^m = (\mathbf{q}_i)^\top \mathbf{a}_m$$



Model: Context-Aware Representation

- **Context-aware bubble**
  - **incorporate information of clicked items related to the target item**

$$\tilde{\mathbf{c}}_m = \sum_{k=1}^{m} \gamma_{km} \mathbf{q}_{i_k}, \quad \text{where } \gamma_{km} = \sigma\left(\frac{(\mathbf{W}_K^3 \mathbf{h}_k)^\top (\mathbf{W}_Q^3 \mathbf{q}_i)}{\sqrt{d}}\right) \quad \tilde{\mathbf{r}}_u^m = Softplus(\mathbf{W}_\mathrm{N}^3 [\tilde{\mathbf{c}}_m \| \mathbf{q}_i] + \mathbf{b}_N^3)$$

- **Estimate with bubble embedding and context-aware state**

inherent interests from observed sequence

+

relations between historical behaviors and target items

$$\tilde{\mathbf{p}}_m = \tilde{\mathbf{c}}_m + \delta(\mathbf{q}_i - \tilde{c}_m) \odot \tilde{\mathbf{r}}_m$$
$$\hat{y}_{ui}^m = (\mathbf{q}_{i_t})^\top \mathbf{a}_m + (\mathbf{q}_{i_t})^\top \tilde{\mathbf{p}}_m$$

Model Optimization: Supervised Learning

- The model estimate the probability with the relevance score

$$P(i|\mathcal{T}_u^m) = \sigma(\hat{y}_{ui}^m)$$

- Adopt Bayesian Personalized Ranking as objective

$$\mathcal{L} = \sum_{u \in \mathcal{U}} \sum_{m=1}^{n_u - 1} \log P(i_{m+1}^u \succcurlyeq \bar{i}_{m+1}^u | \mathcal{T}_u^m)$$

- For the mini-batch data $\{\mathcal{T}_u\}_{u \in \mathcal{U}_b}$

$$\mathcal{L}_{sup} = \sum_{u \in \mathcal{U}_b} \sum_{m=1}^{n_u - 1} \log \sigma(\hat{y}_{u,i_{m+1}^u}^m - \hat{y}_{u,\bar{i}_{m+1}^u}^m)$$

Model Optimization: Contrastive Regularization

- Directly optimize the loss function lead to over-fitting
  - Radius vectors of bubbles tend to be updated radically

- Inspired by contrastive learning
  - Enforce self-consistency within a user sequence
  - Enlarge the mutual information between estimated bubble embedding and historical items
  - Guide the model to ‘look back’

$$\mathcal{L}_{reg} = -\sum_{u \in \mathcal{U}_b} \sum_{m=t+1}^{n_u} \log \frac{\exp(\mathcal{S}(\overline{\mathcal{B}}_u^m, \mathbf{q}_{i^u_{m-t}}))}{\sum_{u' \in \mathcal{U}_b} \exp(\mathcal{S}(\overline{\mathcal{B}}_u^m, \mathbf{q}_{i^{u'}_{m-t}}))}$$

Experiments: Overall Results

Table 1: Comparative results for different methods

Datasets	Metric	POP	BPR-MF	NCF	FPMC	GRURec	GRURec+	Caser	SASRec	TiSASRec	BERT4Rec	DisenRec	Seq2Bubbles	Improv.
Beauty	N@5	0.0241	0.0803	0.0844	0.0921	0.0821	0.1186	0.1054	0.1439	0.1310	0.1585	0.2404	**0.2767**	+13.1%
	H@5	0.0396	0.1219	0.1304	0.1372	0.1321	0.1791	0.1613	0.1929	0.1804	0.2201	0.3225	**0.3508**	+8.0%
	N@10	0.0337	0.1059	0.1132	0.1215	0.1064	0.1448	0.1361	0.1636	0.1566	0.1856	0.2709	**0.2959**	+8.4%
	H@10	0.0755	0.1998	0.2146	0.2415	0.2347	0.2646	0.2593	0.2656	0.2581	0.3029	0.4171	**0.4503**	+7.3%
Steam	N@5	0.0477	0.0744	0.0717	0.0945	0.1370	0.1613	0.1131	0.1727	0.3252	0.1842	0.2863	**0.3566**	+9.7%
	H@5	0.0805	0.1177	0.1203	0.1517	0.2171	0.2391	0.176	0.2559	0.4155	0.2710	0.3986	**0.4384**	+5.5%
	N@10	0.0665	0.1005	0.1026	0.1026	0.1283	0.1802	0.1484	0.2147	0.3557	0.2261	0.3332	**0.3875**	+8.9%
	H@10	0.1389	0.1993	0.2169	0.2551	0.3313	0.3594	0.2870	0.3783	0.5239	0.4013	0.5437	**0.5661**	+4.1%
ML-1m	N@5	0.0416	0.1903	0.1146	0.2885	0.3196	0.3705	0.3832	0.3980	0.4243	0.4454	0.4615	**0.5035**	+9.1%
	H@5	0.0715	0.2866	0.1932	0.4297	0.4673	0.5103	0.5353	0.5434	0.5755	0.5876	0.6025	**0.6351**	+5.4%
	N@10	0.0621	0.2365	0.1640	0.3439	0.3627	0.4064	0.4268	0.4368	0.4641	0.4818	0.5003	**0.5447**	+8.8%
	H@10	0.1358	0.4301	0.3477	0.5946	0.6207	0.6351	0.6692	0.6629	0.7008	0.6970	0.7219	**0.7422**	+2.8%
ML-20m	N@5	0.0511	0.1332	0.0771	0.2239	0.3090	0.3630	0.2538	0.4208	0.5134	0.4967	0.5058	**0.5666**	+10.3%
	H@5	0.0805	0.2128	0.1358	0.3601	0.4657	0.5118	0.3804	0.5727	0.6499	0.6323	0.6528	**0.6931**	+6.1%
	N@10	0.0695	0.1786	0.1271	0.2895	0.3637	0.4087	0.3062	0.4665	0.5440	0.5340	0.5398	**0.6189**	+13.7%
	H@10	0.1378	0.3538	0.2922	0.5201	0.5844	0.6524	0.5427	0.7136	0.7606	0.7473	0.7579	**0.8015**	+5.3%

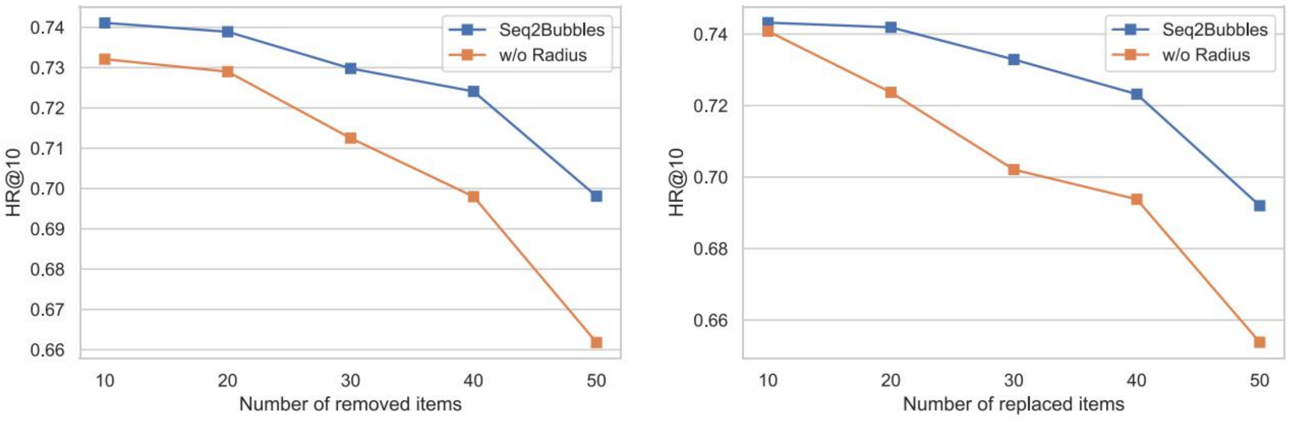
Higher H (HR) and N (NDCG) are better

Experiments: Ablation Study

Table 2: Ablation analysis

Variants	ML-1M		Beauty	
	HR@10	NDCG@10	HR@10	NDCG@10
w/o Contextual	0.731 (-1.4%)	0.536 (-1.5%)	0.422 (-6.2%)	0.276 (-6.4%)
w/o Regularization	0.730 (-1.6%)	0.537 (-1.3%)	0.425 (-5.5%)	0.279 (-5.4%)
w/o Self-Attention	0.621 (-16.3%)	0.483 (-11.2%)	0.352 (-21.7%)	0.183 (-37.9%)
w/o Max Pooling	0.611 (-17.6%)	0.503 (-7.5%)	0.339 (-24.6%)	0.166 (-43.7%)
Default	**0.742**	**0.544**	**0.450**	**0.295**

- Comparison with the simplified version that replace the bubble embedding by point embedding

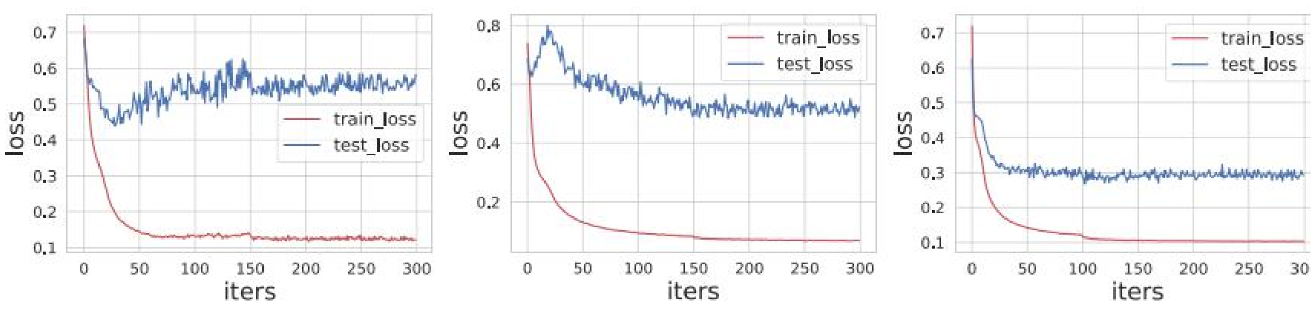


(a) Removing items. (b) Replacing items.

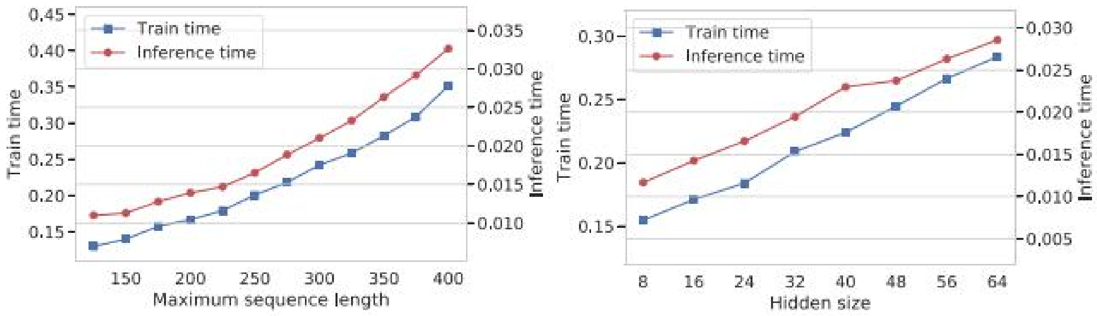
Experiments: Robustness and Scalability

❑ Further discussions:

- The regularization term helps to alleviate over-fitting
- The training time scales linearly w.r.t. sequence length and hidden size



(a) w/o reg. (b) w/ reg. ($t = 10$). (c) w/ reg. ($t = 1$).



(a) Maximum sequence length. (b) Hidden size.

Conclusions

❑ **Our contributions:**

- Methodology: propose a new representation model for distributions of user interests with multi-modality and heterogeneous concentration
- Techniques: design an efficient distance computing scheme of new embedding and devise a self-supervised contrastive to enhance training
- Evaluation: achieve state-of-the-art performance on several benchmarks and conduct ablation studies to thoroughly dissect the effectiveness

Thanks for listening!