NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

TAGON

Temporal Attention Graph Optimized Networks for Sequential Recommendation

Pathak Siddhant (U2023715K)

Supervisor: Assoc Prof Ke Yiping, Kelly
Examiner: Dr. Huang Shell Ying

Introduction – Sequential Recommendation (SR)

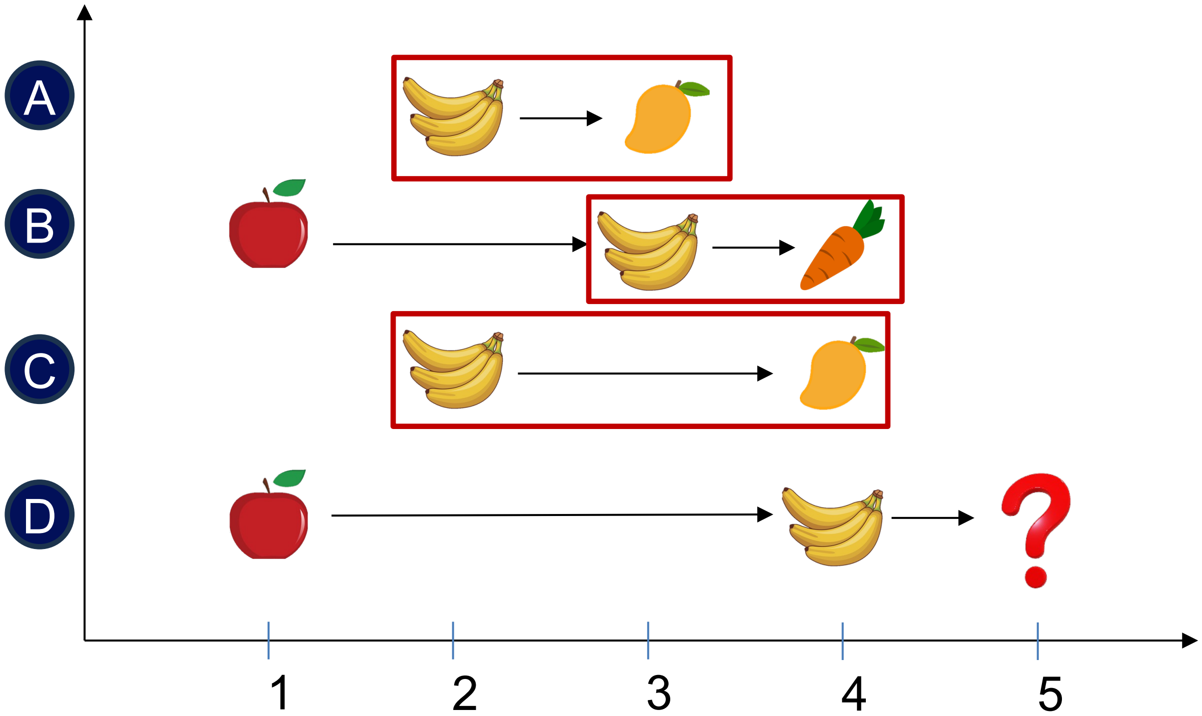
- Personalized recommendations by considering past actions and recent behavior
- SR emphasize interaction order and timing; capture immediate and short-term user interest besides global trend
- Traditional recommendation treat all historical interactions with equal relevance; provides static suggestions based on overall data

Problem Formulation for Sequential Recommendation

Given a user u and their sequence of interactions with items $S = [(\mathrm{i}_1\ \mathrm{t}_1), (\mathrm{i}_2\ \mathrm{t}_2), \ldots (\mathrm{i}_n\ \mathrm{t}_n)]$, such systems aim to maximize the probability $\mathbb{P}(i|(u,S))$ of recommending the next item i that the user is most likely to engage with:

$$\boxed{\mathrm{argmax}_{i\in I}\ \mathbb{P}(i|(u,S))}$$

Introduction – Sequential Recommendation (SR) - Example



Problem Statement

Problem

Traditional recommendation systems often neglect the temporal dynamics of user-item interactions, leading to outdated or irrelevant suggestions.

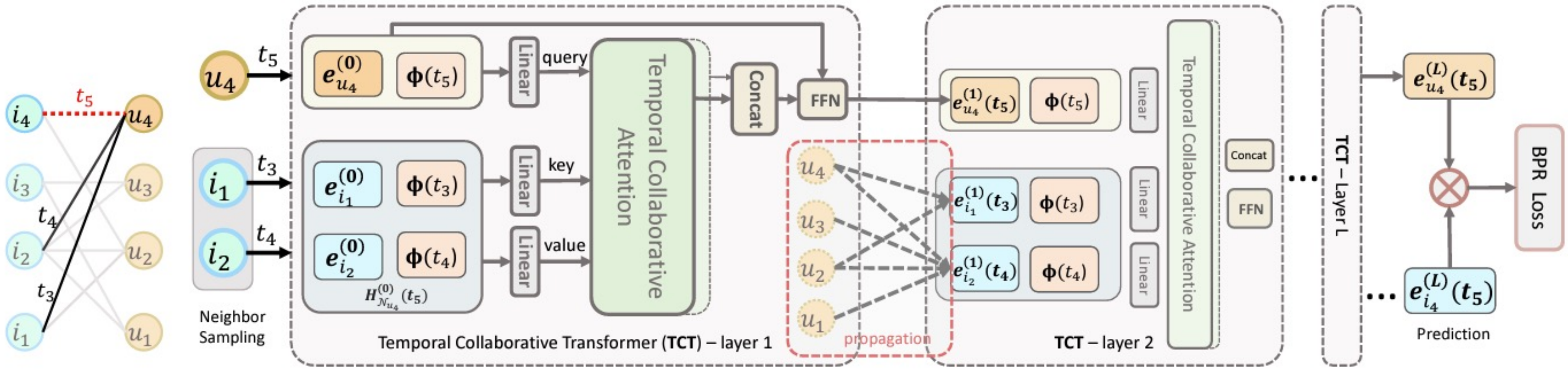
Claim

TAGON address this by integrating temporal dynamics, improving recommendation relevance and timeliness, combining best of graph-based and attention-based learning mechanisms.

- Introduces temporal dynamics into recommendation systems
- Graph Neural Networks enhanced with a temporal attention mechanism
- Adapts dynamically to both sequential and temporal user interactions
- Promises enhanced user engagement through intelligent, context-aware suggestions

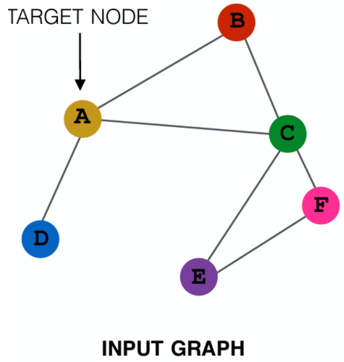
Inspiration behind TAGON - TGSRec

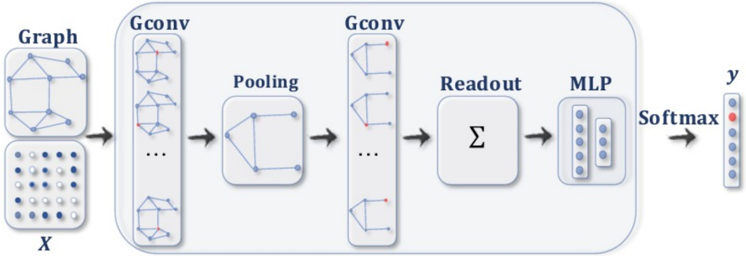
- Developed by Fan et al., 2021
- Introduced a way to fuse different modes of information – interactions + temporal data
- Models time as an encoding mechanism instead of node feature
- Blends together Graph Neural Networks with Attention-based Learning
- Leverages only cross-attention for item and user embedding based refinement
- Does not focus on the relative ordering of items per interaction sequence explicitly

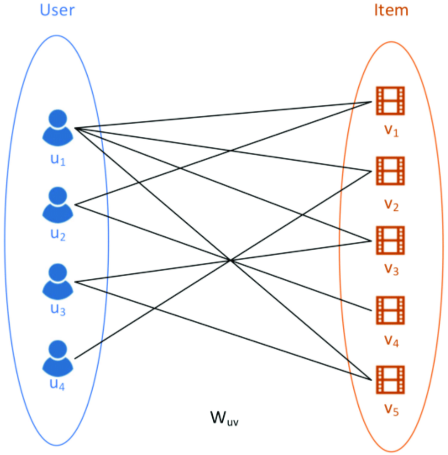


TGSRec architecture [1]

Building Blocks – Graph Neural Networks (GNNs)





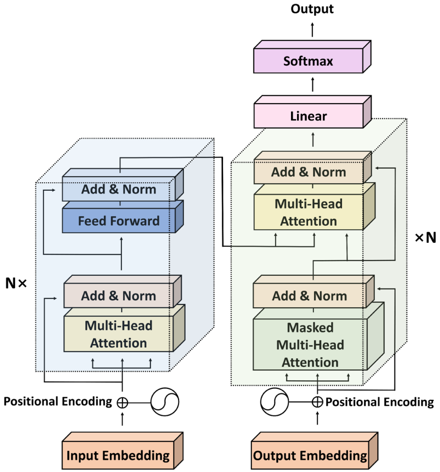


Graph Neural Networks in action [2], [3]

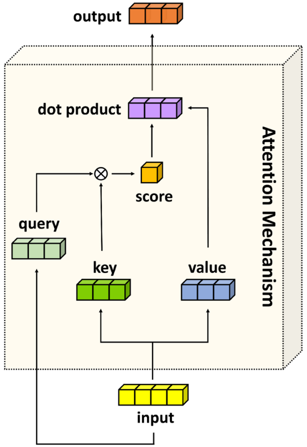
- Models connections between nodes in graph-based data
- Uses message passing for information propagation across nodes
- Learns complex patterns and global graph properties
- Enhances node representations through iterative aggregation and update functions

Building Blocks – Transformers and Attention

- Consists of encoder and decoder with multiple layers
- Utilizes positional encodings to maintain sequence order awareness
- Applies dot-product attention for dynamic weighting of sequence



Transformer architecture [4]

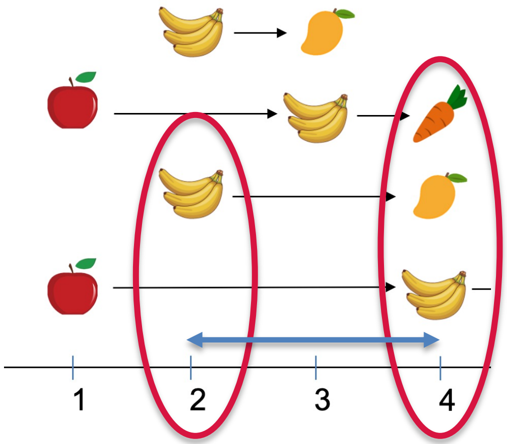


Attention Mechanism [4]

- Computes weights to emphasize relevant parts of data
- Allows models to focus dynamically during processing
- Utilized in networks to enhance feature aggregation

Building Blocks – Sequential Time Data

- Encode time intervals between interactions into vectors
- Utilize a time-encoding function to map timestamps
- Used as an alternative to the traditional positional encoding for attention mechanisms



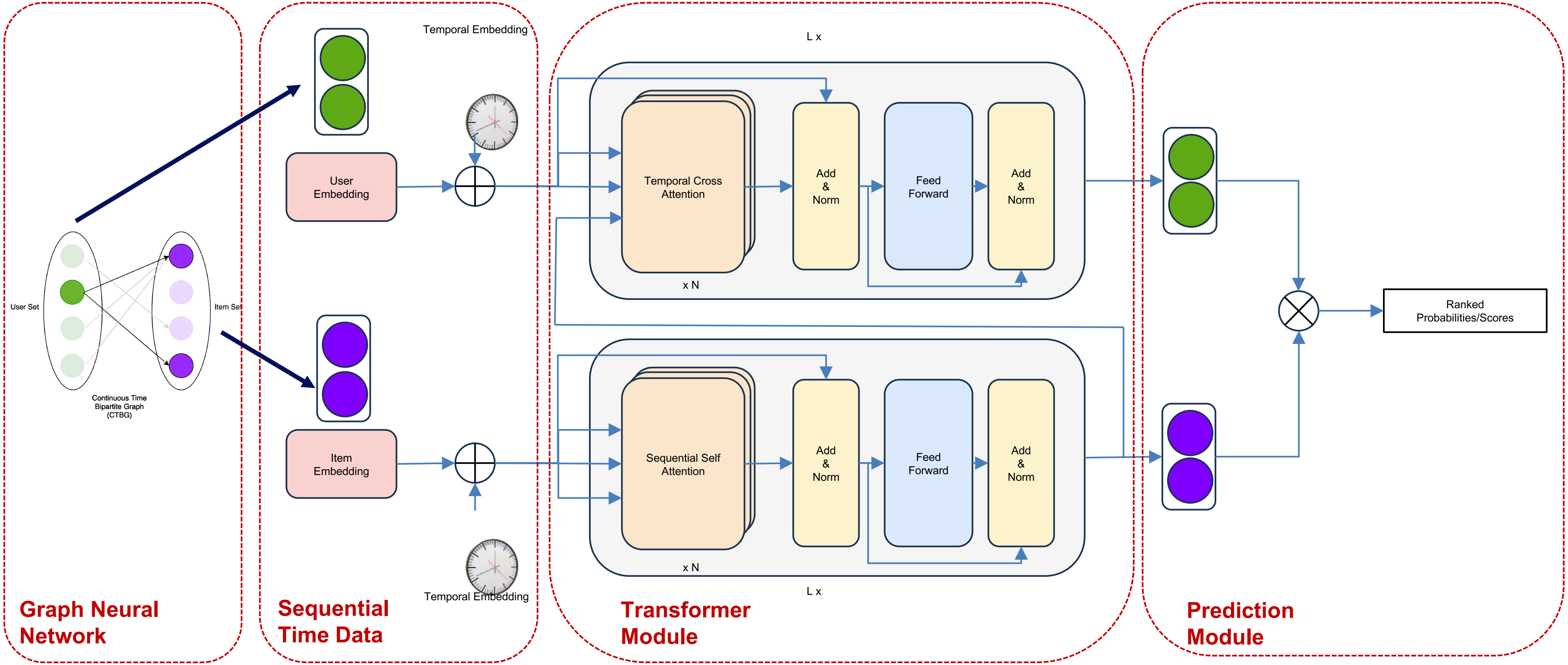
Bochner's Theorem for Timestamp-based Encoding

For a user u and interactions $(\mathrm{i}_1\ \mathrm{t}_1)$ and $(\mathrm{i}_2\ \mathrm{t}_2)$, the time impact is defined as a function of the elapsed time between these events. Formally, it is expressed as:

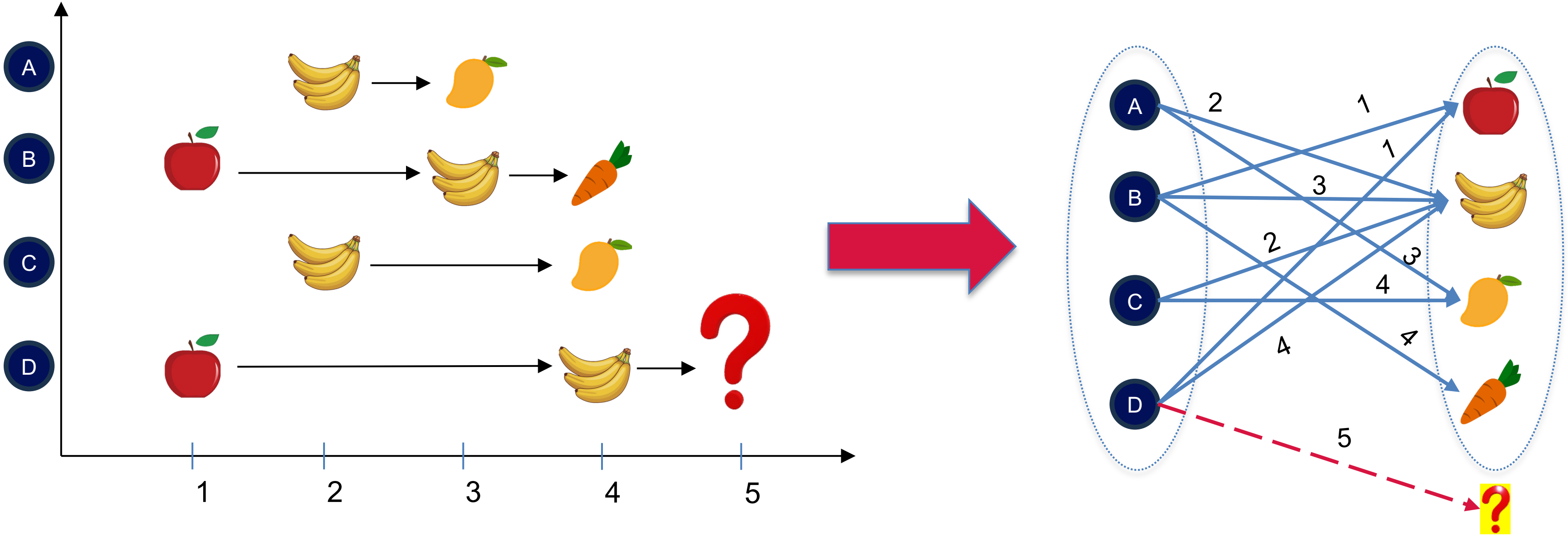
$$\psi(\mathrm{t}_2 - \mathrm{t}_1) = \phi(\mathrm{t}_2) \cdot \phi(\mathrm{t}_1)$$

where $\phi(t)$ is further defined using the Bochner's Theorem, and will be used as the temporal encoding for time t.

Proposed Architecture of TAGON

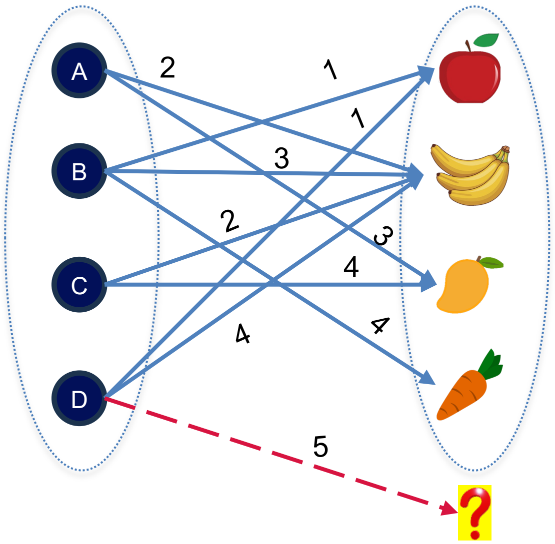


Proposed Architecture of TAGON – Coming back to the example



Continuous Time Bipartite Graph (CTBG)

Proposed Architecture of TAGON – Coming back to the example



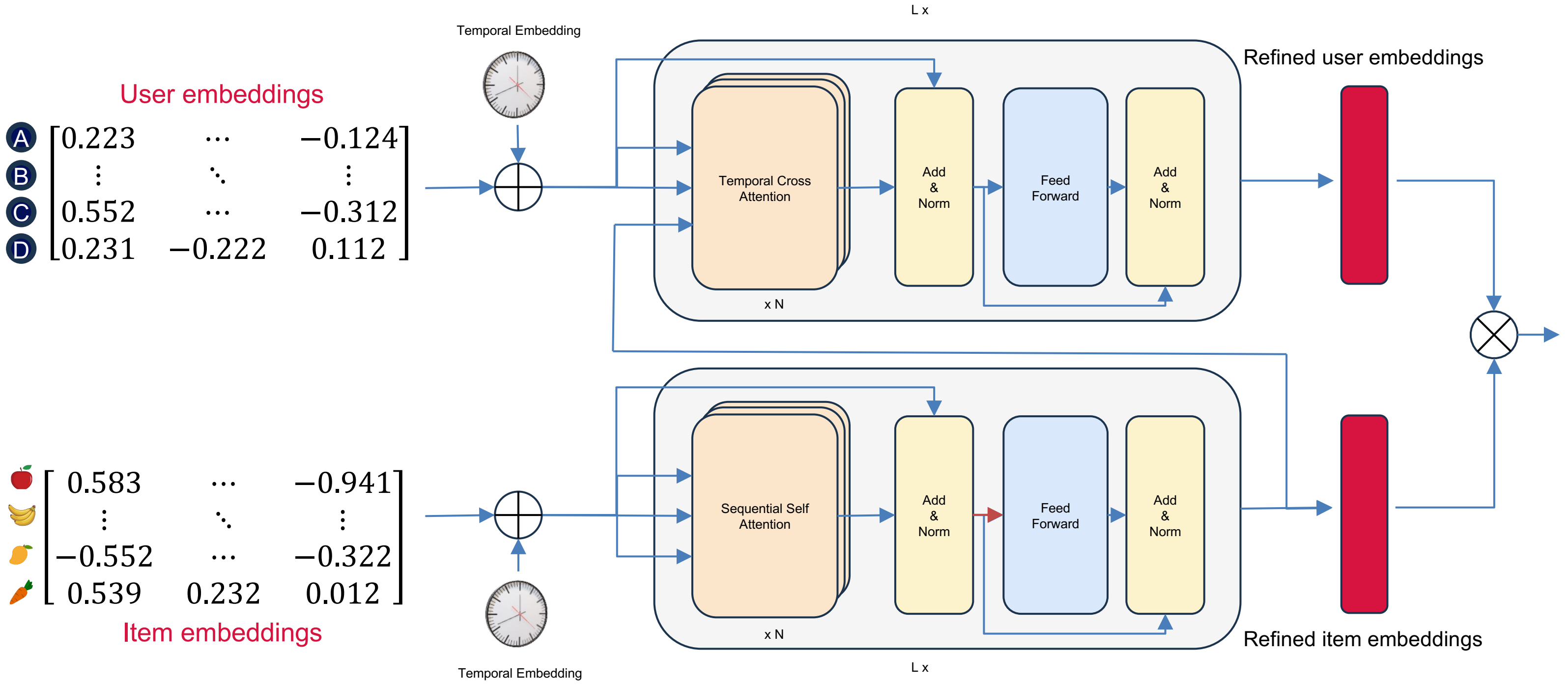
$$
\begin{array}{c}
A \\ B \\ C \\ D
\end{array}
\begin{bmatrix}
0.223 & \cdots & -0.124 \\
\vdots & \ddots & \vdots \\
0.552 & \cdots & -0.312 \\
0.231 & -0.222 & 0.112
\end{bmatrix}
$$

User embeddings

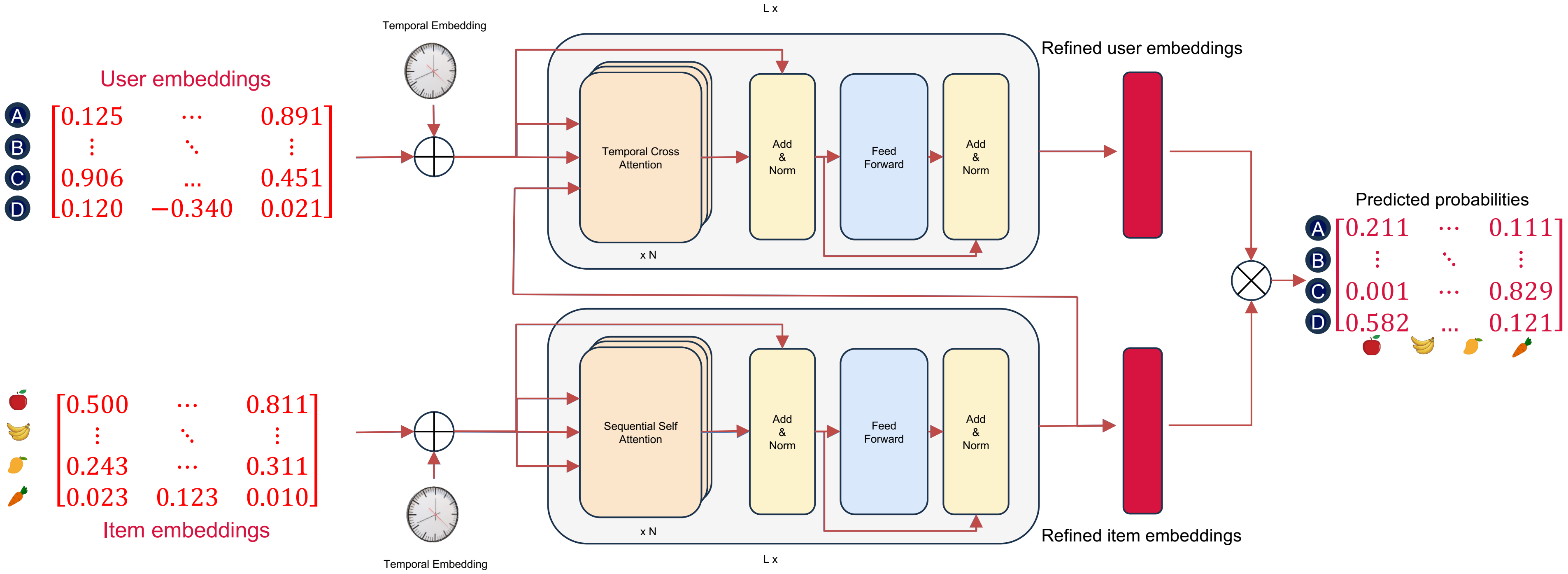
$$
\begin{bmatrix}
0.583 & \cdots & -0.941 \\
\vdots & \ddots & \vdots \\
-0.552 & \cdots & -0.322 \\
0.539 & 0.232 & 0.012
\end{bmatrix}
$$

Item embeddings

Proposed Architecture of TAGON – Coming back to the example



Proposed Architecture of TAGON – Coming back to the example



Experiment Design – Research Questions (RQs)

Research Question 1 (RQ1)

Does the proposed architecture enhance the quality of recommendations?

Research Question 2 (RQ2)

What impact does the introduction of our transformer layer have on our model's performance?

Research Question 3 (RQ3)

In what ways do various encoding techniques influence our model's effectiveness?

Research Question 4 (RQ4)

Is the performance of our model augmented by the implementation of a unifying attention mechanism?

Research Question 5 (RQ5)

How robust is our model to varying hyperparameter settings?

Experiment Design – Datasets, metrics chosen

Datasets

- Amazon Review Datasets
  - Baby
  - Toys and Games
  - Tools and Home Improvement
  - Digital Music
- Movielens (100K variant)

Dataset	Toys	Baby	Tools	Music	ML-100K
# Users	17,946	17,739	15,920	4,652	943
# Items	11,639	6,876	10,043	3,051	1,682
# Edges	154,793	146,775	127,784	54,932	48,569
# Train	134,632	128,833	107,684	51,765	80,003
# Validation	11,283	10,191	10,847	2,183	1,516
# Test	8,878	7,751	9,523	984	1,344
Sparsity	99.93%	99.88%	99.92%	99.62%	93.70%

Metrics for Evaluation

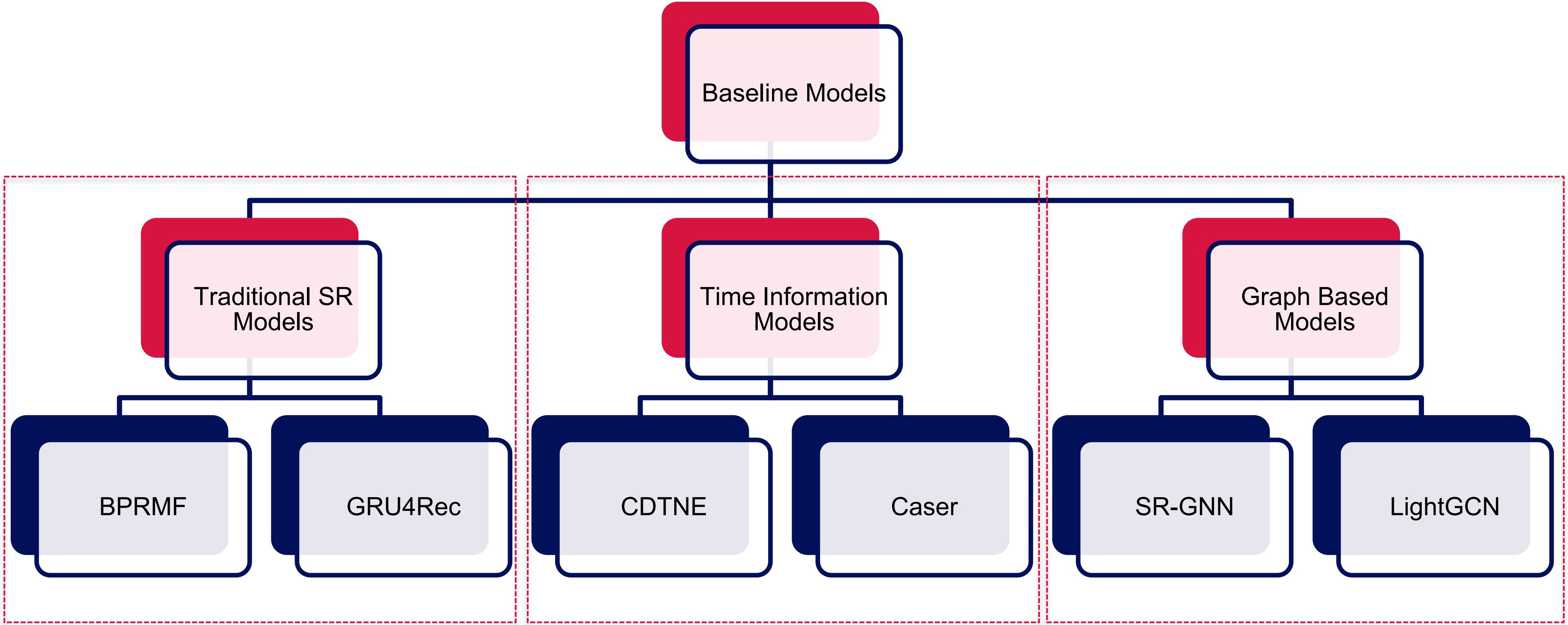
$$Recall@K = \frac{\#\ relevant\ items\ in\ K}{Total\ \#\ of\ relevant\ items}$$

$$MRR = \frac{1}{U}\sum_{u=1}^{U} rank_i$$

Loss Function

$$\mathcal{L} = \underbrace{\sum_{(u,i,j,t)\in O_T} -\log\left(\sigma\left(r(u,i,t) - r(u,j,t)\right)\right)}_{\text{Bayesian Pairwise Ranking Loss}} + \underbrace{\lambda\|\Theta\|^2}_{L_2\text{ regularization}}$$

Experiment Design – Baseline Models chosen



Results – Performance Evaluation (RQ1)

Research Question 1

Does the proposed architecture enhance the quality of recommendations?

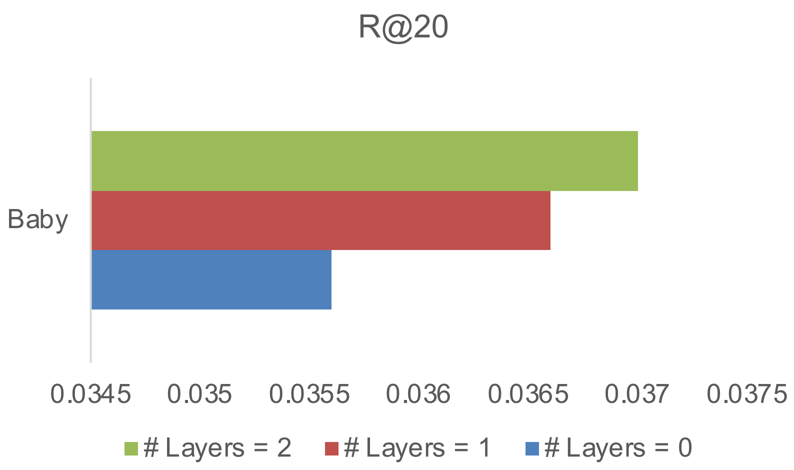
	Model	Toys		Baby		Music		Tools		ML-100K	
		R@10	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10	MRR
Traditional SR	BPR	0.0021	0.0024	0.0028	0.0019	0.0122	0.0057	0.0023	0.0026	0.0461	0.0213
	GRU4Rec	0.0274	0.0201	0.0036	0.0028	0.0495	0.0540	0.0048	0.0051	0.0554	0.0938
Time Information	Caser	0.0302	0.0082	0.0077	0.0071	0.0183	0.0106	0.0077	0.0068	0.0246	0.0147
	CDTNE	0.0016	0.0025	0.0218	0.0157	0.0071	0.0037	0.0186	0.0191	0.0350	0.0162
Graph Based	LightGCN	0.0016	0.0018	0.0036	0.0024	0.0142	0.0064	0.0021	0.0023	0.0565	0.0252
	SR-GNN	0.0020	0.0018	0.0030	0.0024	0.0051	0.0028	0.0051	0.0028	0.0045	0.0012
	My Model	0.0310	0.0204	0.0248	0.0167	0.0522	0.0617	0.0192	0.0202	0.0691	0.0982
	Improvement (%)	**2.53%**	**1.49%**	**11.99%**	**5.93%**	**5.24%**	**12.42%**	**3.31%**	**5.35%**	**18.25%**	**4.49%**

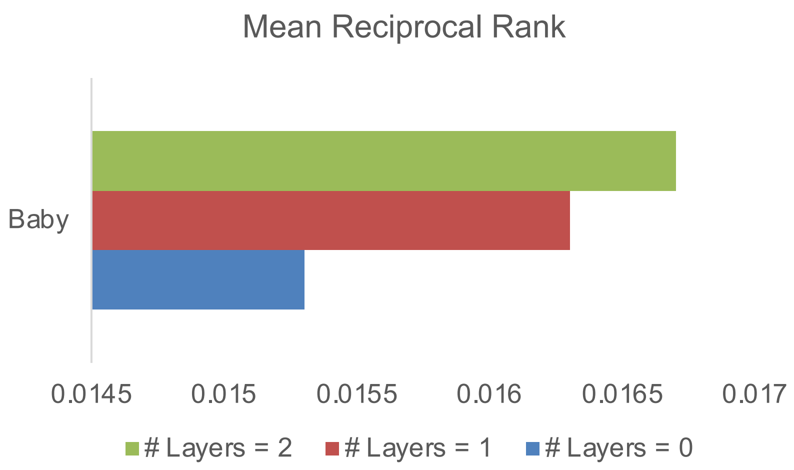
TGSRec was not benchmarked due to lack of time and resources

Results – Ablation Study (RQ2)

Research Question 2

What impact does the introduction of our transformer layer have on our model's performance?

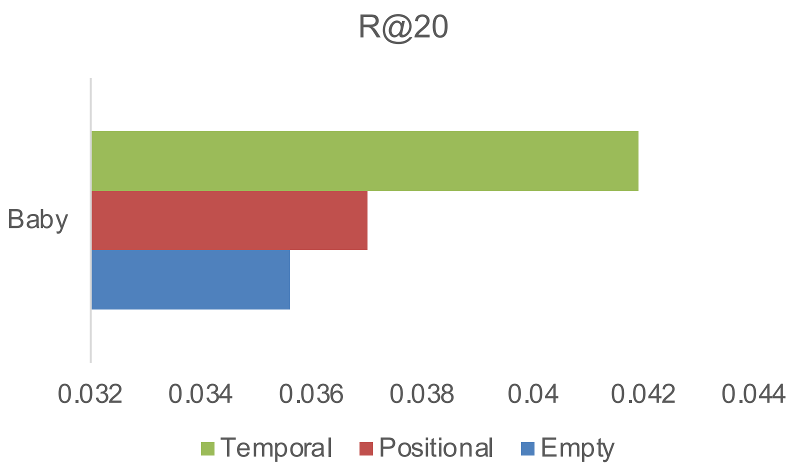


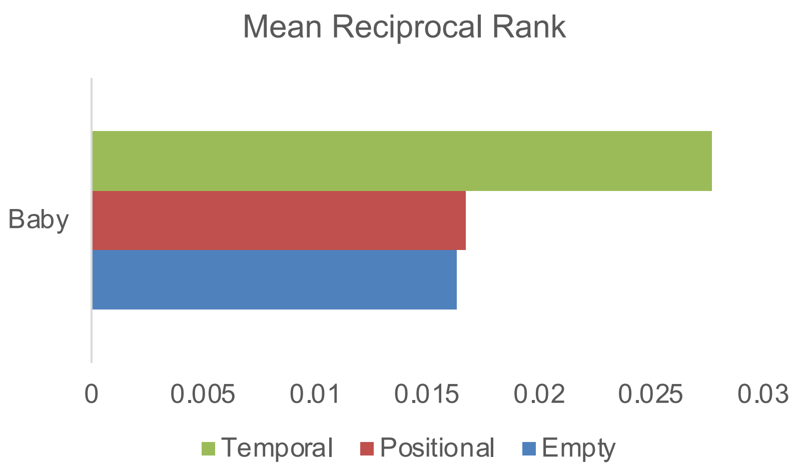


Results – Ablation Study (RQ3)

Research Question 3

In what ways do various encoding techniques influence our model's effectiveness?

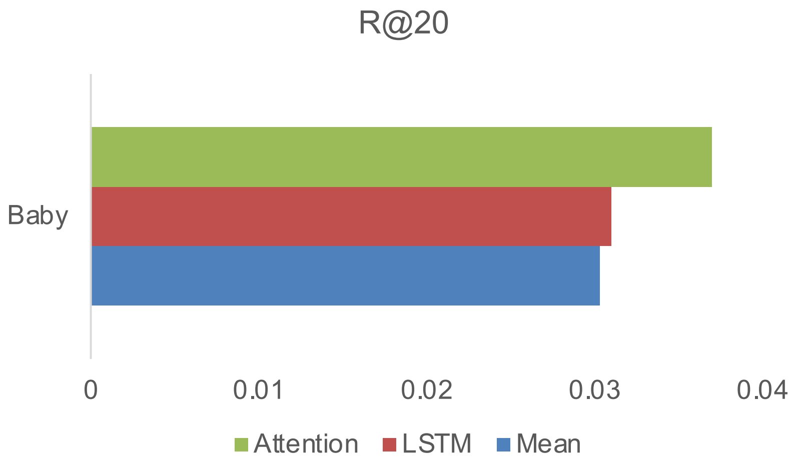


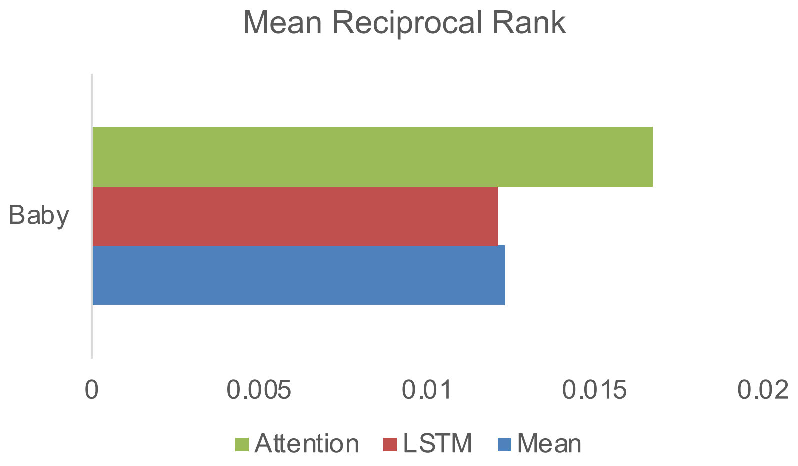


Results – Ablation Study (RQ4)

Research Question 4

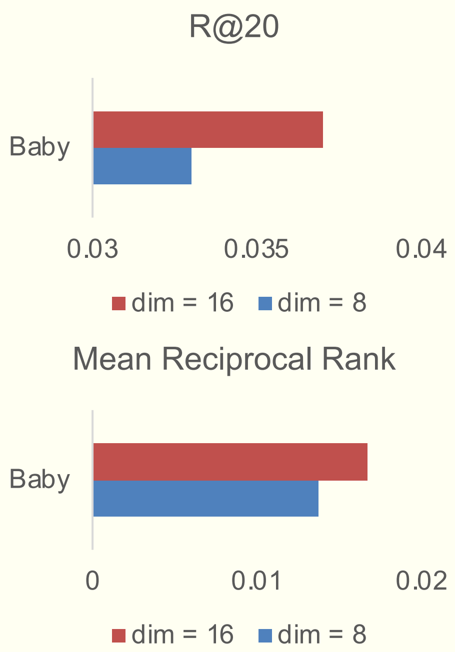
Is the performance of our model augmented by the implementation of a unifying attention mechanism?



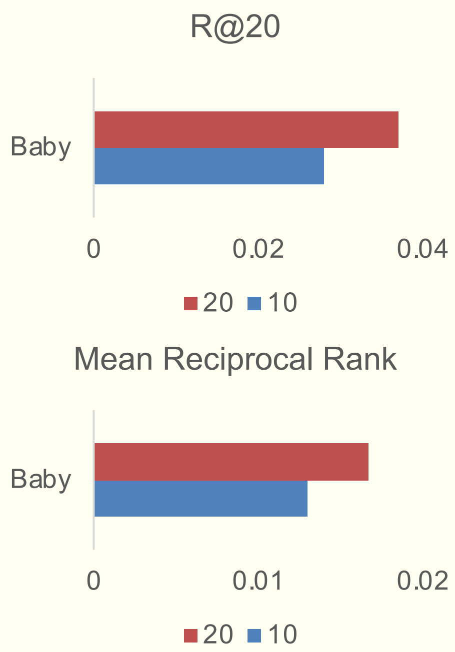


Results – Hyperparameter Tuning (RQ5)

Embedding Dimensions



Degree of neighbor sampling



Conclusion and Future Works

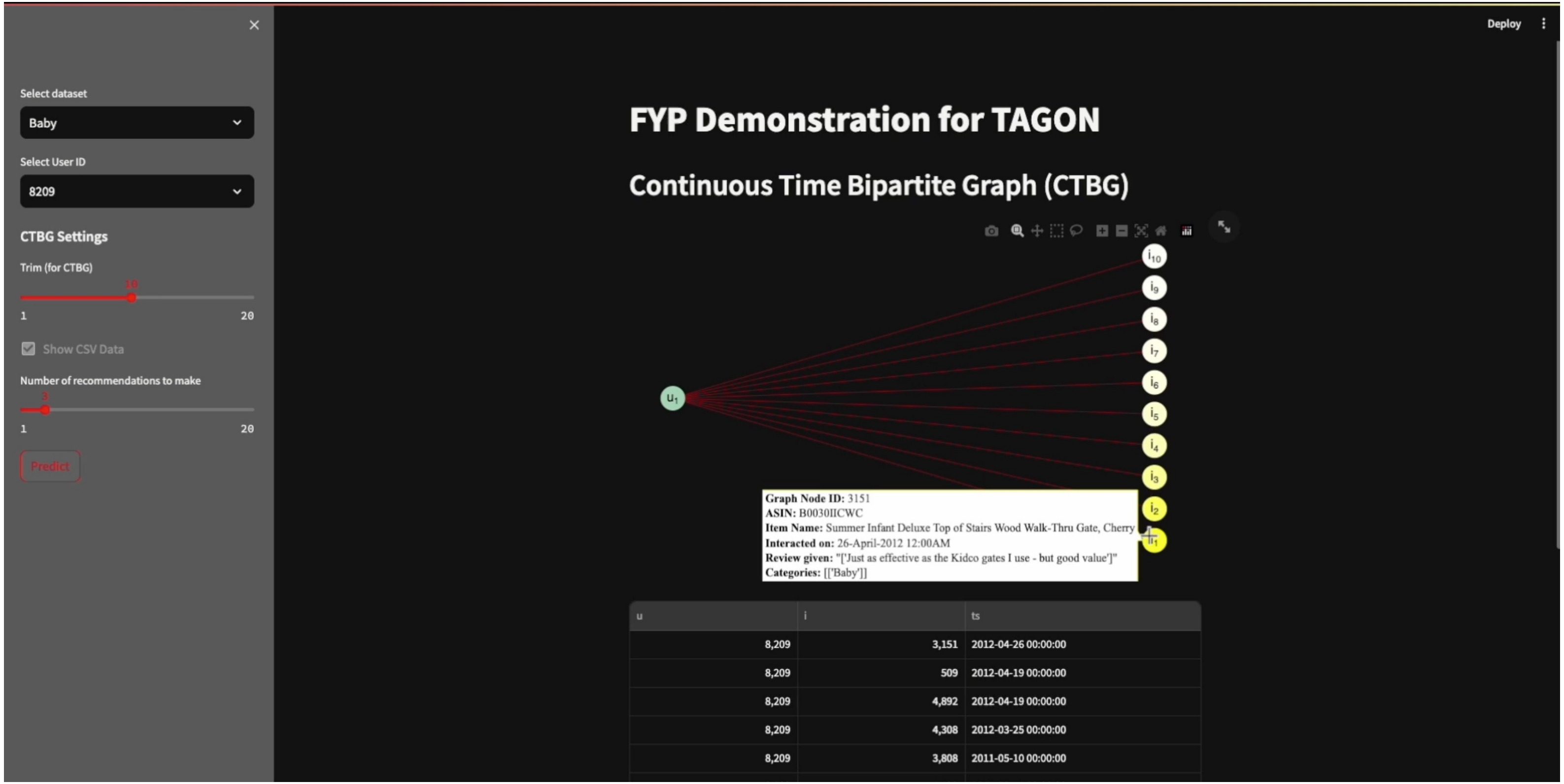
Limitations	Future Work
Lacks explainability, transparency	Explore attention weights from model output
Separation of long, short-term interests	Extract information using subgraphs
Long training and evaluation time periods	New sampling distribution for embeddings
Not capturing local patterns efficiently	Explore use of Pointwise CNNs as FFNs

Conclusion

- Introduced a novel framework to capture and leverage complex user-item interactions
- Infused time as source of information for nodes, to account for dynamic nature of users, items
- Utilized user-item graphs, with attention mechanism to prioritize relevant interactions, ensuring context relevance
- Conducted meticulous experimentation against six models from new and old domains of SR
- Demonstrated significant improvements in both accuracy and relevance of recommendation

Code is publicly available at www.github.com/siddhantpathakk/tagon

Demonstration for TAGON



Please view video here: https://shorturl.at/glzF3

References

1. Ziwei Fan et al. “Continuous-Time Sequential Recommendation with Temporal Graph Collaborative Transformer”. en. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Virtual Event Queensland Australia: ACM, Oct. 2021, pp. 433–442.

2. Leskovec, J. *Graph neural networks*. *CS422W: Machine Learning with Graphs (Stanford University)*.

3. Leskovec, J. *How expressive are GNNs? CS422W: Machine Learning with Graphs (Stanford University)*.

4. Amer, H. (n.d.). *What are attention mechanisms in the context of neural networks, especially transformers?*. Quora. https://www.quora.com/How-do-neural-networks-learn-What-are-attention-mechanisms-in-the-context-of-neural-networks-especially-transformers