Déjà vu: A Contextualized Temporal Attention Mechanism for Sequential Recommendation

By Jibang Wu, Renqin Cai, Hongning Wang

Present by: Tianyang Chen, Veronique Wang, Zetao Wang, Chenlin Liu

Part 1

Introduction

- **C**ontextualized **T**emporal **A**ttention Mechanism (CTA)
    - weigh historical actions’ influence on **what action it is**, **when** and **how the action took place**.

Challenges

The influence patterns from different segments of history reflect user interests in different ways.

Temporal Segment

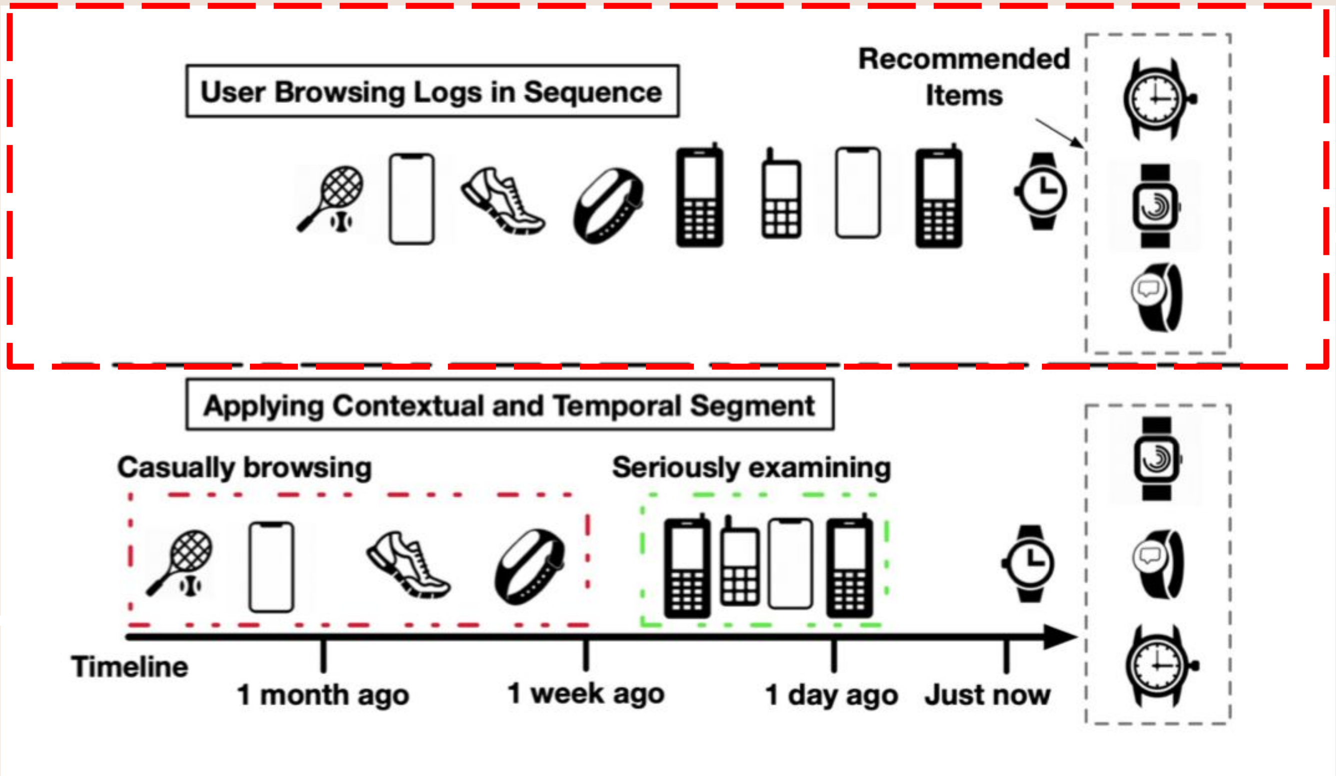
- Distant user history: sparse yet crucial information of user preferences in general
- Recent user history: closely represent the user intention in near future.

Contextual Segment

- In general: user browsing log appears to be heterogeneous
- Certain point: the user concentrate on a small subset of homogeneous items

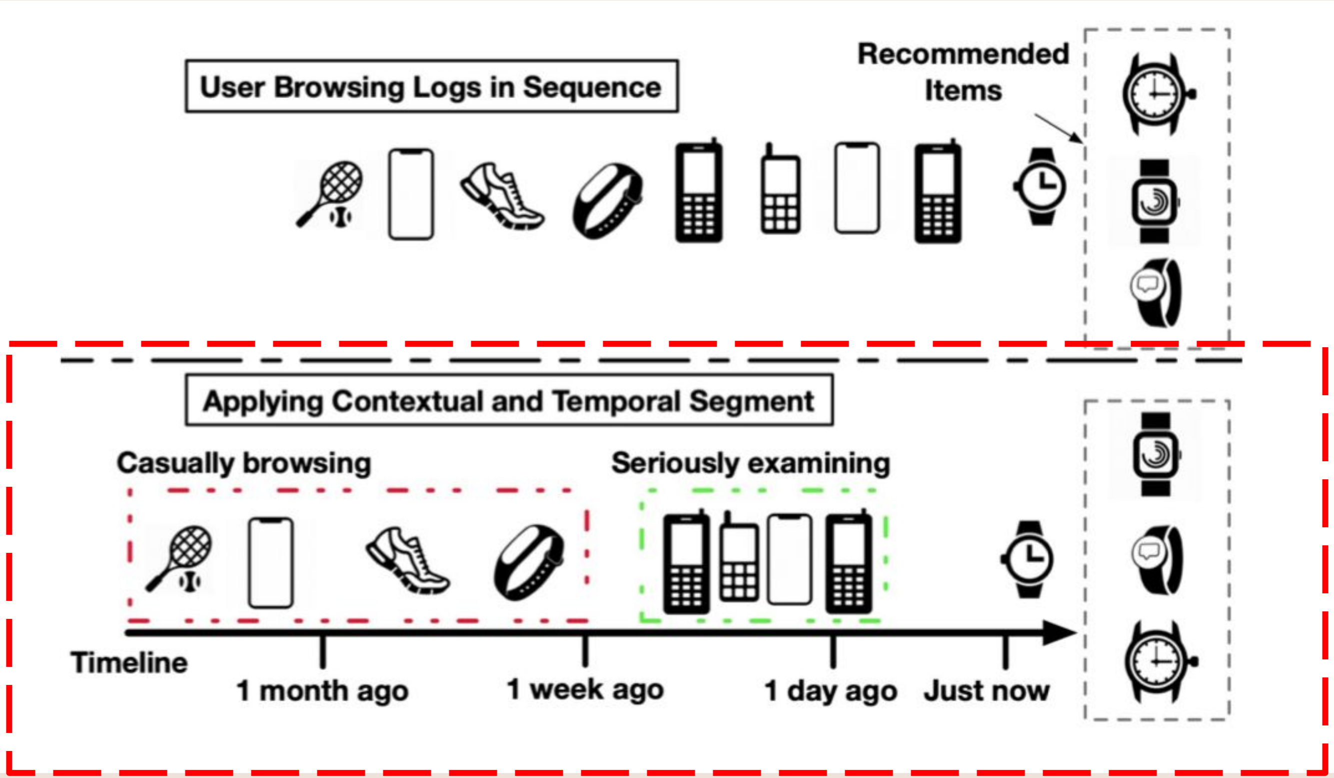
Goal: to capture and connect these different signals from each part of history

Example



User Browsing Logs in Sequence
Recommended Items
Applying Contextual and Temporal Segment
Casually browsing
Seriously examining
Timeline
1 month ago
1 week ago
1 day ago
Just now

Example



User Browsing Logs in Sequence
Recommended Items
Applying Contextual and Temporal Segment
Casually browsing
Seriously examining
Timeline
1 month ago
1 week ago
1 day ago
Just now

Related Works

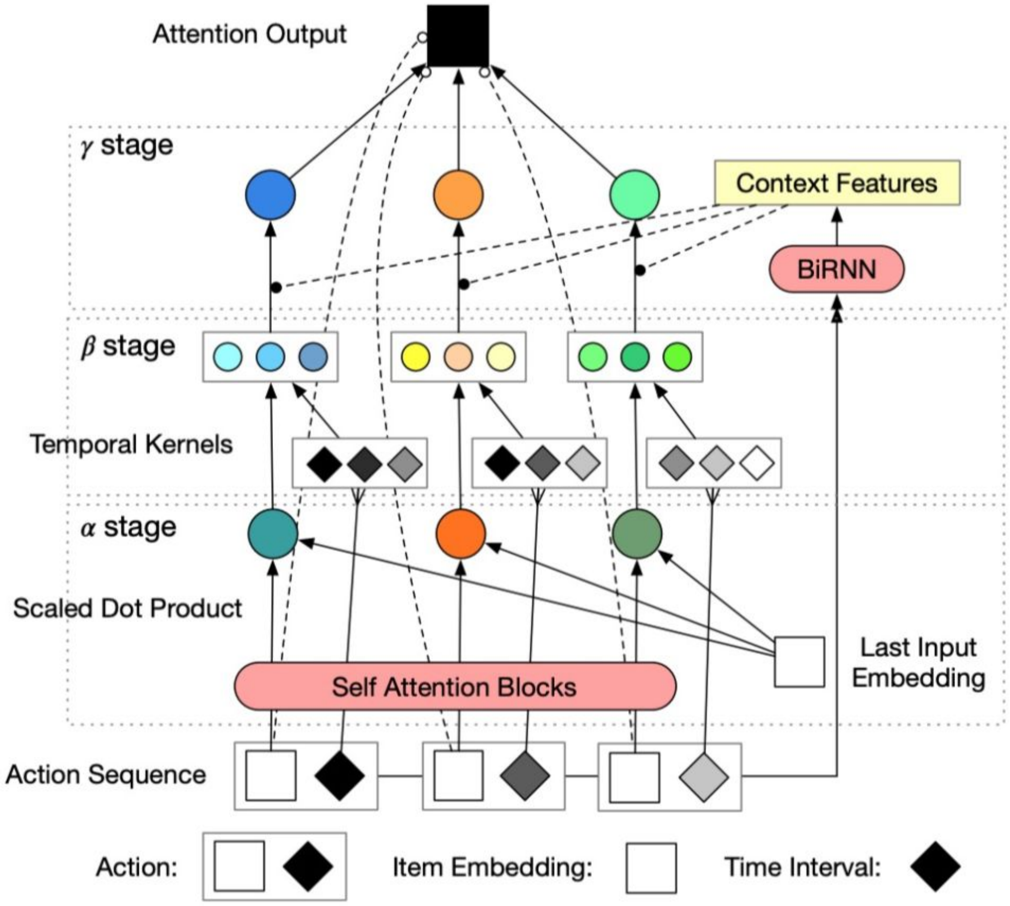
Sequential Recommendation

- Time-based sessions: RNN
    - Long Short-Term Memory
    - Gated Recurrent Units
- Convolutional Neural Networks (CNN), Memory Network, and Attention Models
- Self-attention mechanism

Temporal Recommendation

- Model separately the long-term static and short-term dynamic user preference
- Matrix factorization
- Time series analysis
    - Hawkes process based algorithms model

Model


Attention Output
γ stage
Context Features
BiRNN
β stage
Temporal Kernels
α stage
Scaled Dot Product
Last Input
Embedding
Self Attention Blocks
Action Sequence
Action:
Item Embedding:
Time Interval:

Model Setup

Consider the sequential recommendation problem with temporal information:

Denote the item space as $\mathcal{V}$ of size N, and the user space as $\mathcal{U}$ of size U.

Given a set of user behavior sequences from users:

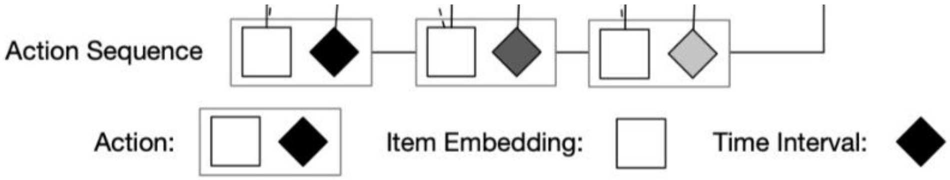
$$\mathcal{C} = \{\{(t_i^u, s_i^u)\}_{i \in \mathcal{I}_u}\}_{u \in \mathcal{U}}$$

Model Setup

$$\mathcal{C} = \{\{(t_i^u, s_i^u)\}_{i \in \mathcal{I}_u}\}_{u \in \mathcal{U}}$$

1. C is consisted of a series of item-time tuples, where t_j^u is the timestamp, and s_i^u is the item accessed by the user.

2. Timestamp is represented by the *real-value scalar*. Item is represented as an *embedding vector*.



Model Setup

Attributes: Sweetness Sourness Softness

Apple: [0.5, 0.1, 0.4]

Banana: [0.7,0.1,0.2]

Pineapple: [0.2,0.5,0.3]

Grape: [0.8,0.1,0.1]

Coconut: [0.9,0.0.1]

Shopping History: 3 bananas, 1 coconut

Next Recommendation?

Model Setup

Heuristic Method: Majority Voting

Shopping History: 3 bananas, 1 coconut

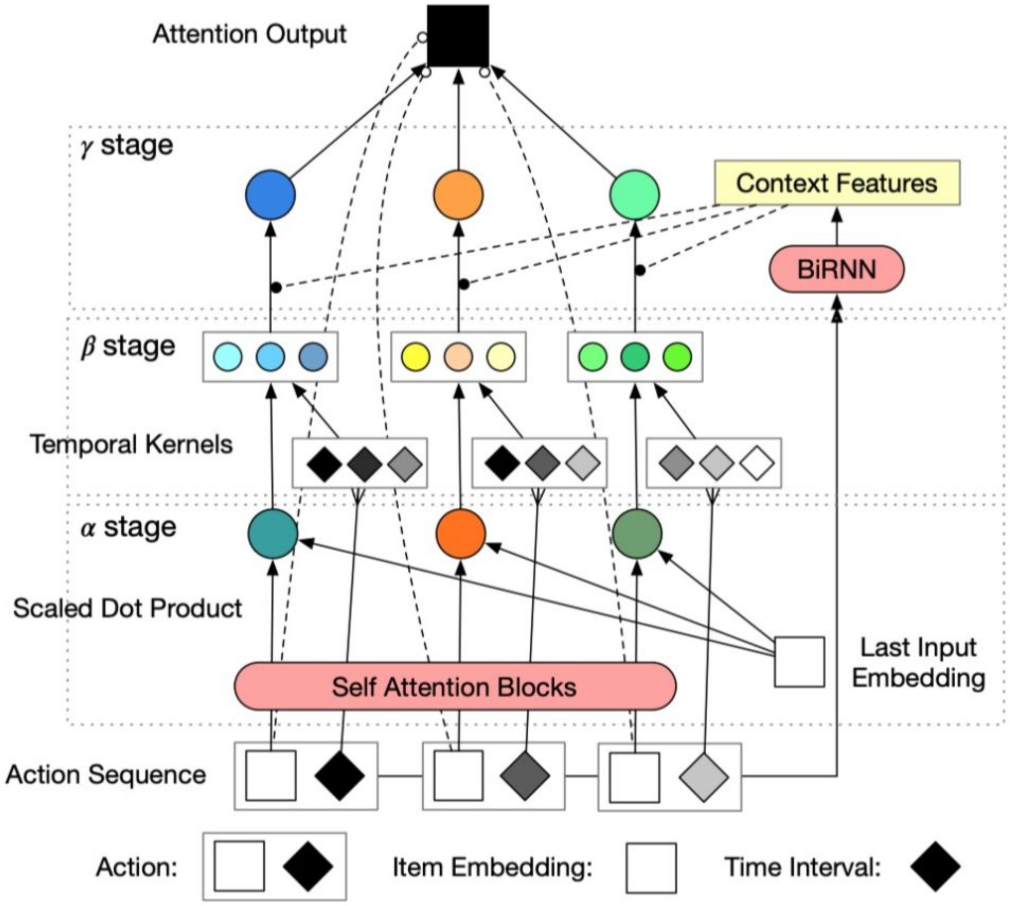
(3 x [0.7,0.1,0.2] + [0.9,0.0.1]) / 4

= **[0.75, 0.075, 0.175]** -> vector C

Dot product

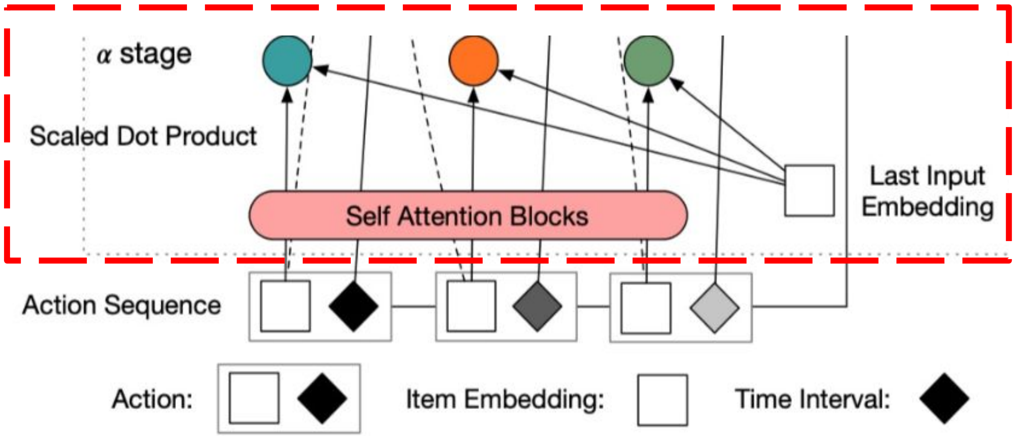
Apple: [0.5, 0.1, 0.4]	0.0425
Banana: [0.7,0.1,0.2]	0.0576
Pineapple: [0.2,0.5,0.3]	0.24
Grape: [0.8,0.1,0.1]	0.625
Coconut: [0.9,0.0.1]	0.6925

Model



Attention Output
γ stage
Context Features
BiRNN
β stage
Temporal Kernels
α stage
Scaled Dot Product
Last Input Embedding
Self Attention Blocks
Action Sequence
Action:
Item Embedding:
Time Interval:

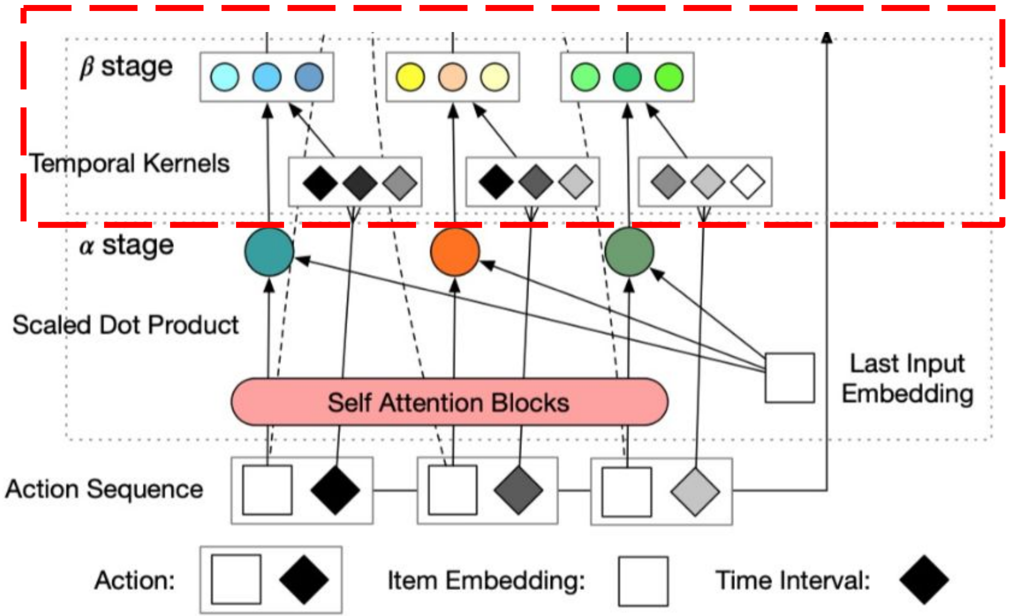
Model – Alpha Stage



- **Goal:** *content-importance score* based on item embedding, $\boldsymbol{X}$
- **Approach:** Encoder mode of the self-attentive model
  - Transform input items embedding $\boldsymbol{X}$ into last layer hidden representation $\boldsymbol{H}$.
  - Computes *bilinear attention product*

$$\boldsymbol{\alpha} = \mathrm{softmax}\left(\frac{\boldsymbol{H} W_\alpha x^T}{\sqrt{d}}\right)$$

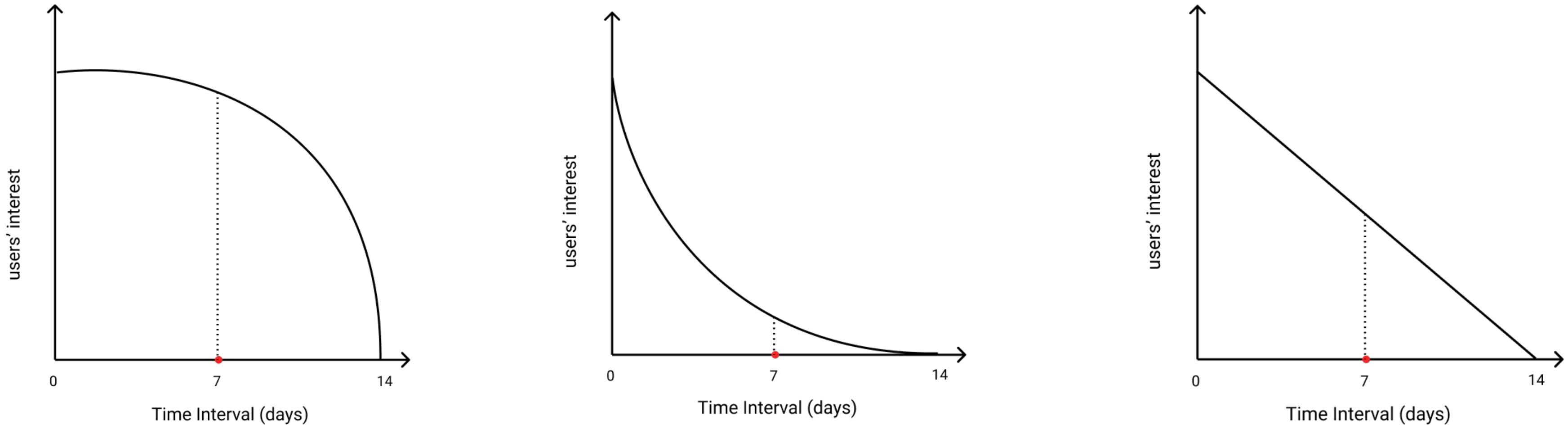
Model – Beta Stage



- **Goal:** to determine the past events' influence based on their *temporal gaps* from the current moment of recommendation.

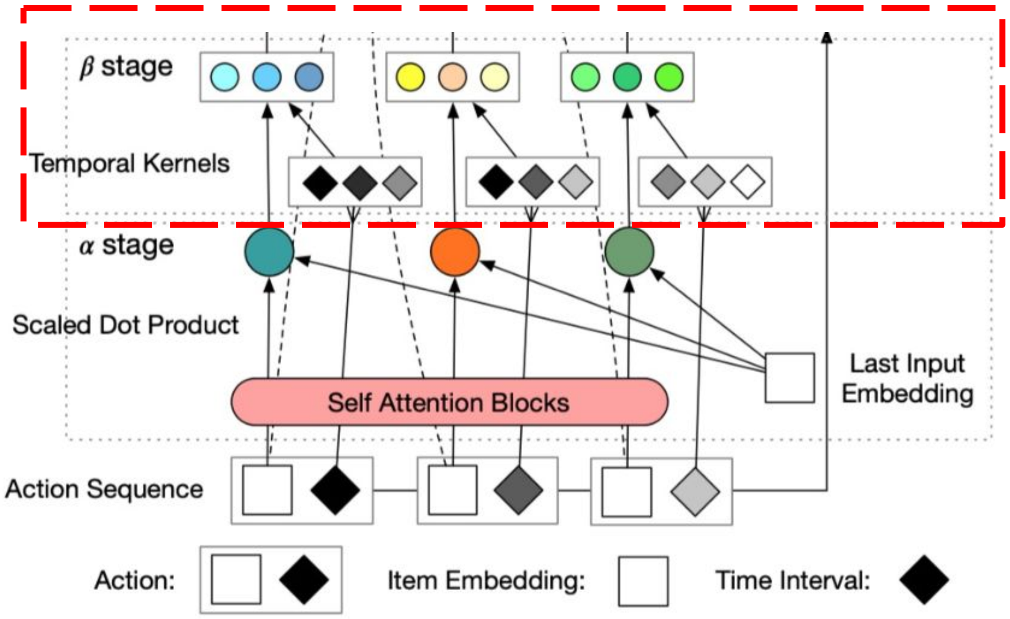
Model – Beta Stage

Possible kernel functions:



and so much more...

Model – Beta Stage



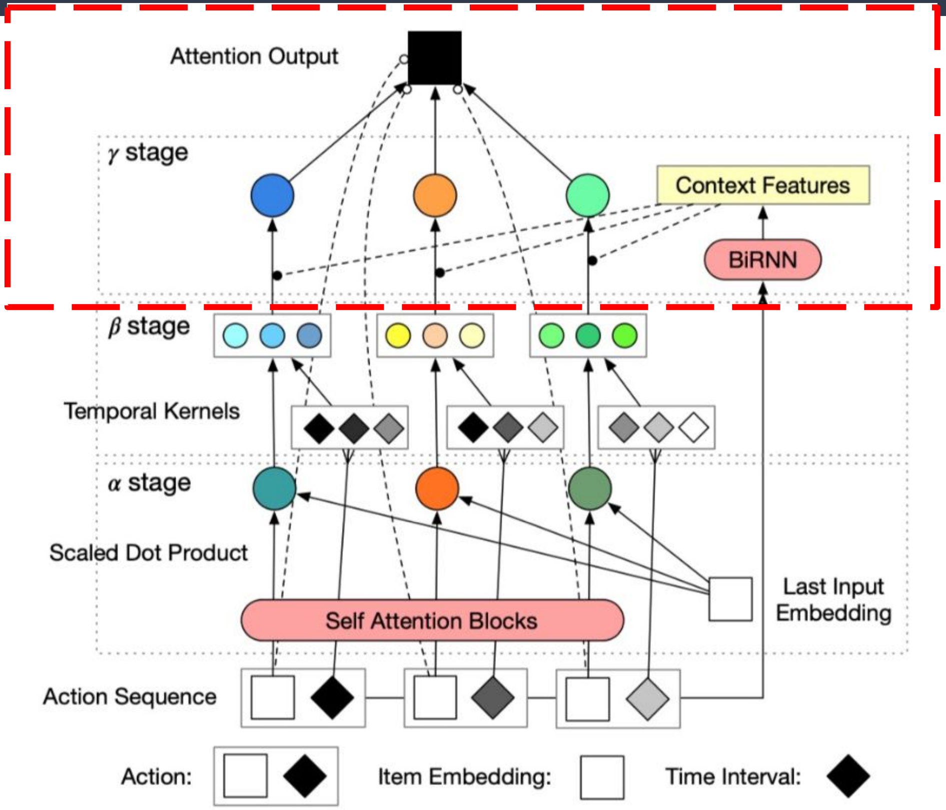
- We pick some kernel function of different shapes: $\phi(\cdot) : \mathbb{R}^L \to \mathbb{R}^L$.
- To model the various temporal dynamics, this stage learns a set of *K* different temporal kernel functions:

$$\left\{\phi(\cdot)^1, \dots, \phi(\cdot)^K\right\}$$

and then transform T into a vector of *temporal importance scores*:

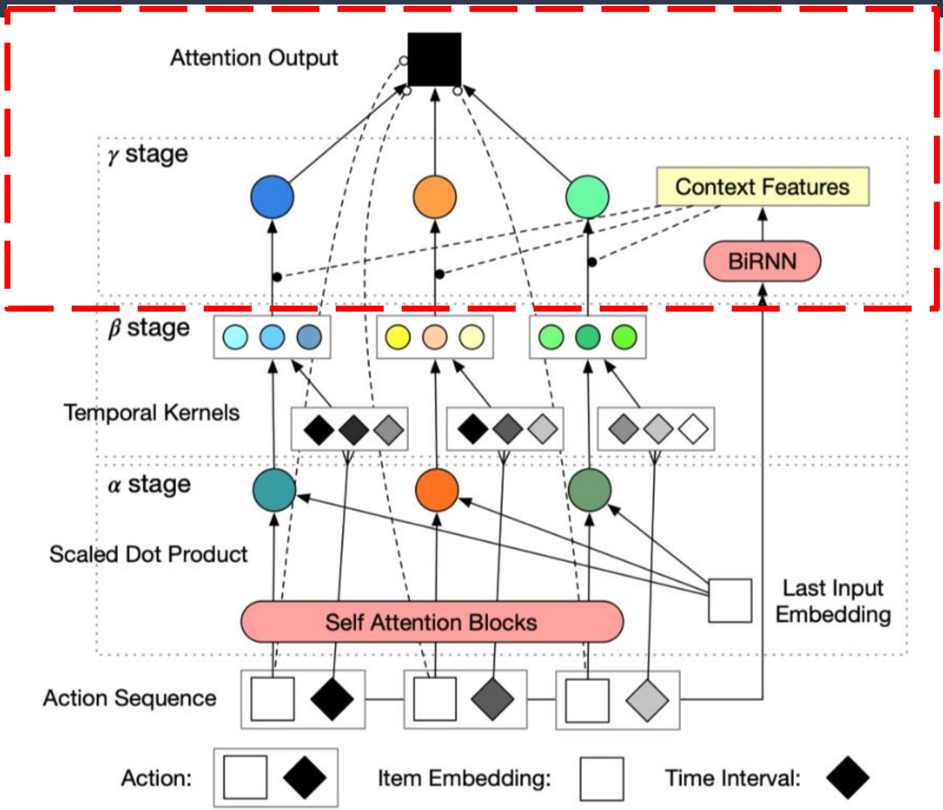
$$\boldsymbol{\beta} = \left[\phi^1(T), \dots, \phi^K(T)\right]$$

Model – Gamma Stage



- **Goal:** to fuse the content and temporal influence based on the extracted *context information*

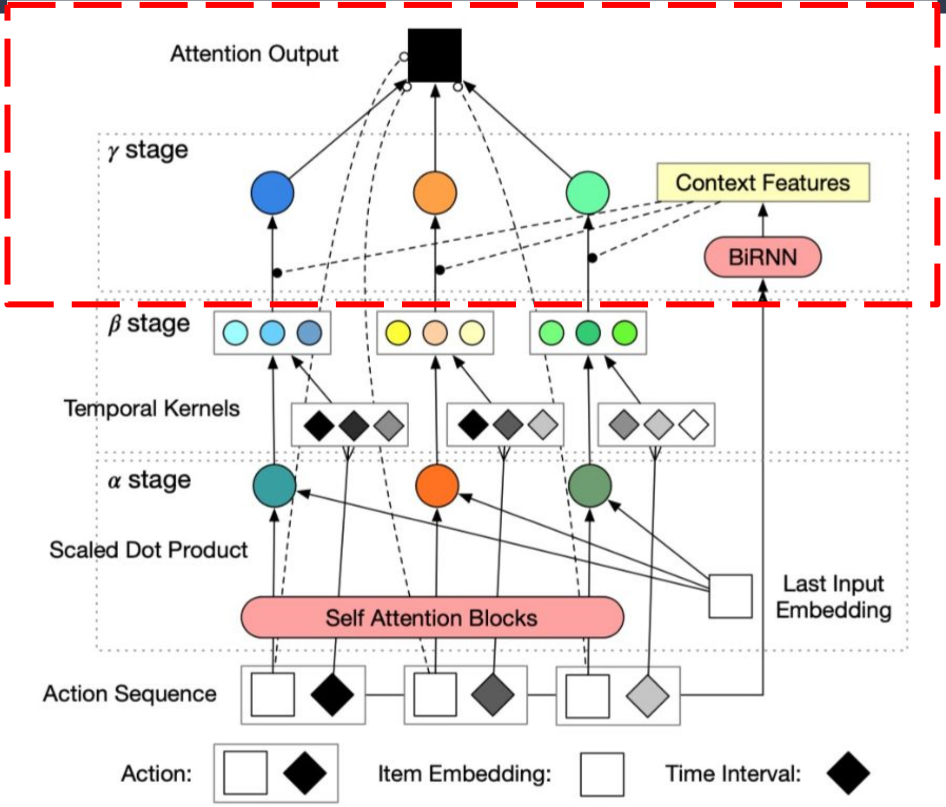
Model – Gamma Stage



1. Extract Context Information by extract the *bidirectional recurrent hidden state* of the input item embedding $\boldsymbol{X}$ as the context feature vector:

$$\boldsymbol{C} = \text{Bi-RNN}(X) \oplus \boxed{C_{\text{attr}}}$$

Model – Gamma Stage

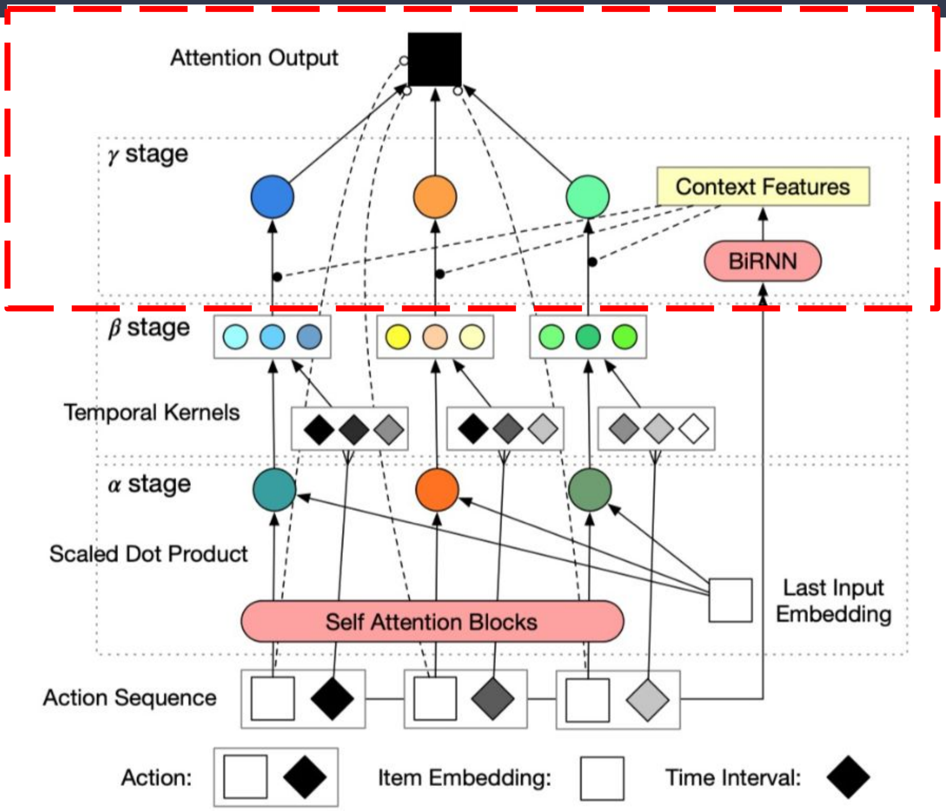


2. Compute the *contextualized temporal influence score* from beta weighted by the probability distribution based on this vector

$$\boldsymbol{P}(\cdot|\boldsymbol{C}) = \mathrm{softmax}\left(F^{\gamma}(\boldsymbol{C})\right)$$

$$\boldsymbol{\beta}^{c} = \boldsymbol{\beta} \cdot \boldsymbol{P}(\cdot|\boldsymbol{C})$$

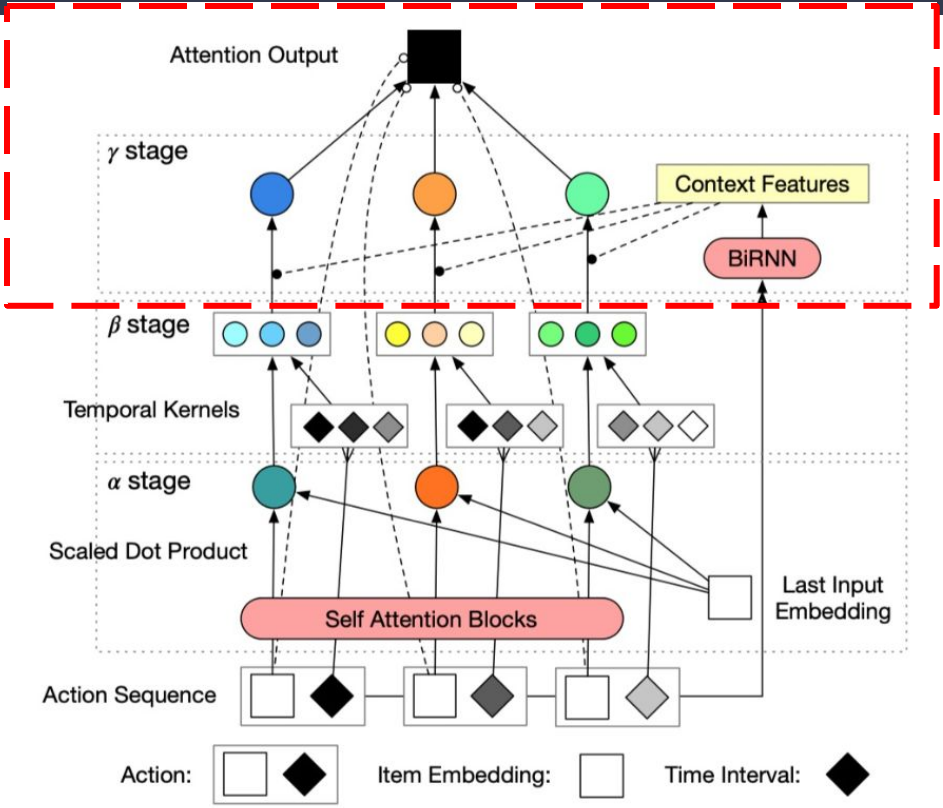
Model – Gamma Stage



3. Infuse the *importance score* from content **α** and *contextualized temporal factors* $\boldsymbol{\beta}^c$

$$\boldsymbol{\gamma} = \mathrm{softmax}\left(\boldsymbol{\alpha}\boldsymbol{\beta}^c\right)$$

Model – Gamma Stage



4. Compute *attention output* for item similarity scores with each item embedding E_i then recommend items with the highest scores r_i

$$\hat{x} = F^{\text{out}}(\gamma^T \cdot \boldsymbol{X}) \quad \forall i \in \mathcal{V}, r_i = E_i \cdot \hat{x}$$

Part 4

Experiment

Dataset

- User Behavior
  - user interactions on commercial products from an e-commerce website
- XING
  - user actions on job postings from a professional social network site
- Attributes
  - user ID, item ID, action timestamp and interaction type (click, favor, purchase, etc)

Experiment Setup

Baseline method

- Heuristics Methods
  - Global Popularity (Pop), Sequence Popularity (S-Pop), First Order Markov Model (Markov)
- Session-based Models
  - Session based Recurrent Neural network (GRU4Rec), Hierarchical Recurrent Neural network (HRNN)
- Temporal Models
  - Long- and Short-term Hawkes Process (LSHP)
- Sequential Models
  - Self-attentive Sequential Recommendation (SASRec), Multi-temporal-range Mixture Model (M3R)

Experiment Setup

Evaluation metrics

- Recall at K
  - Reports the percentage of times that the groundtruth relevant item
  - Ranked within the top K list of retrieved items
- MRR at K
  - The mean reciprocal rank is used to evaluate the prediction quality from the predicted ranking of relevant items

Experimental results

- XING dataset features **first order transition**
- UserBehavior dataset features **sequence popularity**
- The model's MRR@5 shows **weak** performance on XING dataset

Dataset	XING		UserBehavior	
Metric	Recall@5	MRR@5	Recall@5	MRR@5
CTA	***0.3217***	0.1849	***0.1611***	***0.0925***
Pop	0.0118	0.0062	0.0026	0.0013
S-Pop	0.2059	0.1202	**0.1093**	**0.0639**
Markov	**0.2834**	**0.2319**	0.0846	0.0534
GRU4Rec	0.2690	0.2008	0.0936	0.0619
HRNN	0.2892	0.2392	0.0940	0.0610
LSHP	0.2173	0.1454	0.1201	0.0792
SASRec	0.2530	0.2254	0.1418	0.0863
M3R	0.2781	***0.2469***	0.1077	0.0689

Performance Analysis

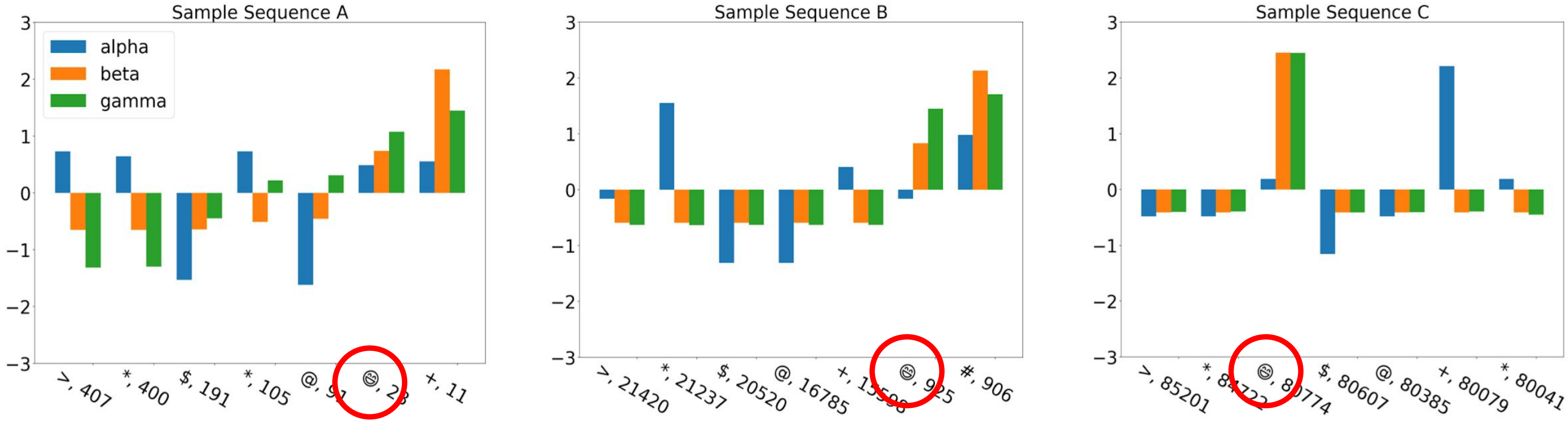


Figure 4: Attention visualization. The blue (left) bar is the content-based importance score α, the orange (middle) bar is the contextualized temporal influence score β^c, the green (right) bar is the combined importance score γ. The figures contains three different sequences selected from the test set of the UserBehavior dataset.

Part 5

Conclusion

Advantages

- Efficacy and Efficiency
- Interpretability
- Customizability

Limitations and Future Works

- Assumptions about the users
- Collaborative Learning

Questions？