(24th Sep)

Paper information:

* title : MEANTIME: Mixture of Attention Mechanisms with Multi-Temporal Embeddings for Sequential Recommendation

* authors : Cho, Sung Min, Eunhyeok Park, and Sungjoo Yoo

* venue : RecSys 2020

* pdf link : http://arxiv.org/abs/2008.08273

* github link : https://github.com/SungMinCho/MEANTIME

* Abstract

Recently, self-attention based models have achieved state-of-the-art performance in sequential recommendation task. Following the custom from language processing, most of these models rely on a simple positional embedding to exploit the sequential nature of the user's history. However, there are some limitations regarding the current approaches. First, sequential recommendation is different from language processing in that timestamp information is available. Previous models have not made good use of it to extract additional contextual information. Second, using a simple embedding scheme can lead to information bottleneck since the same embedding has to represent all possible contextual biases. Third, since previous models use the same positional embedding in each attention head, they can wastefully learn overlapping patterns. To address these limitations, we propose MEANTIME (MixturE of AtteNTIon mechanisms with Multi-temporal Embeddings) which employs multiple types of temporal embeddings designed to capture various patterns from the user's behavior sequence, and an attention structure that fully leverages such diversity. Experiments on real-world data show that our proposed method outperforms current state-of-the-art sequential recommendation methods, and we provide an extensive ablation study to analyze how the model gains from the diverse positional information.

Summary

problems to address

- Previous sequential recommendation models did not utilize timestamp information well, only sequence information. The authors use timestamp information fully with multiple types of temporal embeddings and mixture of attention heads.

key ideas

- TiSASRec [1] introduces Personalized Time Intervals and use Time Interval-Aware Self-Attention.

Each weight coefficient α_{ij} is computed using a softmax function:

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}} \tag{7}$$

And e_{ij} is computed using a compatibility function that considers inputs, relations and positions:

$$e_{ij} = \frac{m_{s_i} W^Q \left(m_{s_j} W^K + r_{ij}^k + p_j^k \right)^T}{\sqrt{d}}, \tag{8}$$

where $W^Q \in \mathbb{R}^{d \times d}$, $W^K \in \mathbb{R}^{d \times d}$ are input projections for a query and key respectively. The scale factor $\sqrt{d}$ is used to avoid large values of the inner product, especially when the dimension is high.

- This paper propose MEANTIME (MixturE of AtteNTIon mechanisms with Multi-temporal Embeddings) model which employs multiple types of temporal embeddings designed to capture various patterns from the user's behavior sequence, and an attention structure that fully leverages such diversity.

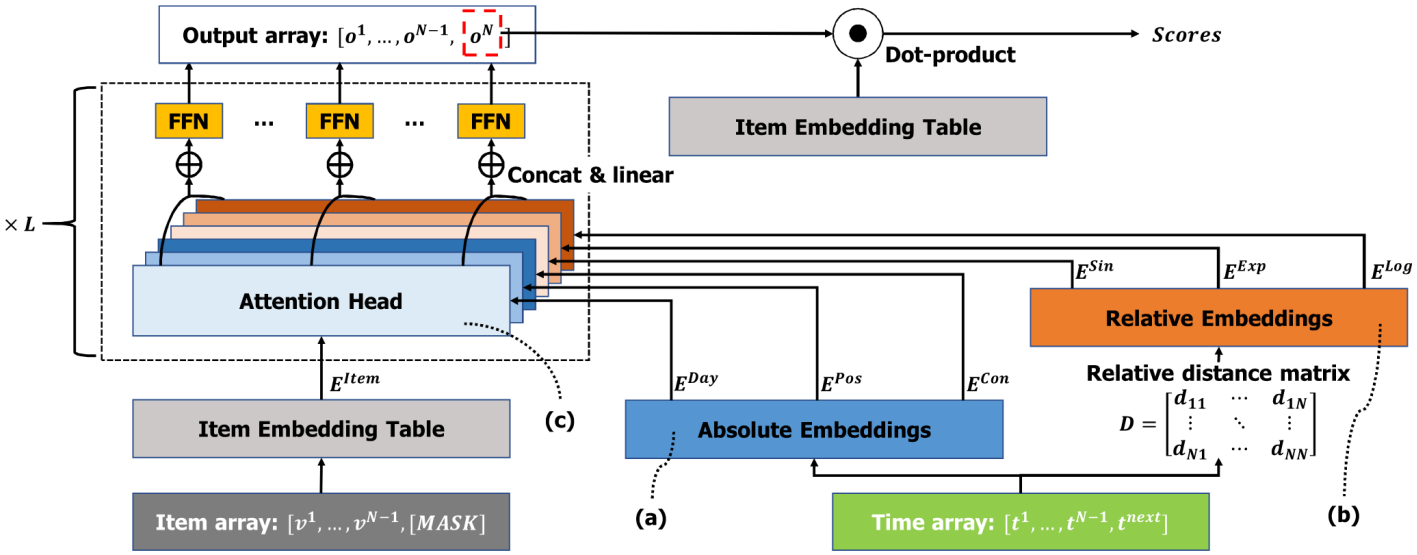

Fig. 1. Model Architecture

3.2 Input Embedding

Our model is based on a fixed-length model with length N as in previous works [10, 19]. Given the history (V_u, T_u) of user u, we feed $\mathbf{v} = [v^u_{|V^u|-N+2}, \ldots, v^u_{|V^u|}, [\text{MASK}]]$ and $\mathbf{t} = [t^u_{|V^u|-N+2}, \ldots, t^u_{|V^u|}, t^u_{next}]$ to the model. Then the model estimates v_{next} as output. If the history is shorter than $N-1$, it is padded by a special token [PAD].

- The authors propose six methods to embed t: three absoulte methods (Figure 1(a), and three relative methods (Figure 1(b)).
  - Absolute embeddings:
    - **Day**-embedding which converts the day of each timestamp into an embedding vector to get E^{Day}
    - **Pos**-embedding is analogous to the learnable positional embedding and converts each position to get E^{Pos}.
    - **Con**-embedding is similar to Pos-embedding, except that all position vectors are shared ($M^C \in R^{1 \times h}$) to remove positional bias. Hence, E^{Con} is a stack of repeated vectors.
  - Relative embeddings:

First, we define a matrix of temporal differences $\mathbf{D} \in \mathbb{R}^{N \times N}$ whose element is defined as $d_{ab} = (\mathbf{t}_a - \mathbf{t}_b)/\tau$, where τ is an adjustable unit time difference. We introduce three encoding functions for **D**. First, **Sin** encoder converts the difference d_{ab} to a hidden vector $\vec{\theta}_{ab} \in \mathbb{R}^{1 \times h}$, by the following equation:

$$\vec{\theta}_{ab,2c} = sin(\frac{d_{ab}}{freq^{\frac{2c}{h}}}) \qquad \vec{\theta}_{ab,2c+1} = cos(\frac{d_{ab}}{freq^{\frac{2c}{h}}}) \tag{1}$$

where $\vec{\theta}_{ab,c}$ is the c^{th} value of the vector $\vec{\theta}_{ab}$ and $freq$ is an adjustable parameter. Likewise, **Exp** encoder and **Log** encoder converts d_{ab} to $\vec{e}_{ab}$ and $\vec{l}_{ab}$ respectively by applying the following equations:

$$\vec{e}_{ab,c} = exp(\frac{-|d_{ab}|}{freq^{\frac{c}{h}}}) \qquad \vec{l}_{ab,c} = log(1 + \frac{|d_{ab}|}{freq^{\frac{c}{h}}}) \tag{2}$$

Stacking these vectors gives us E^{Sin}, E^{Exp} and E^{Log} respectively.

3.3 Self-Attention Structure

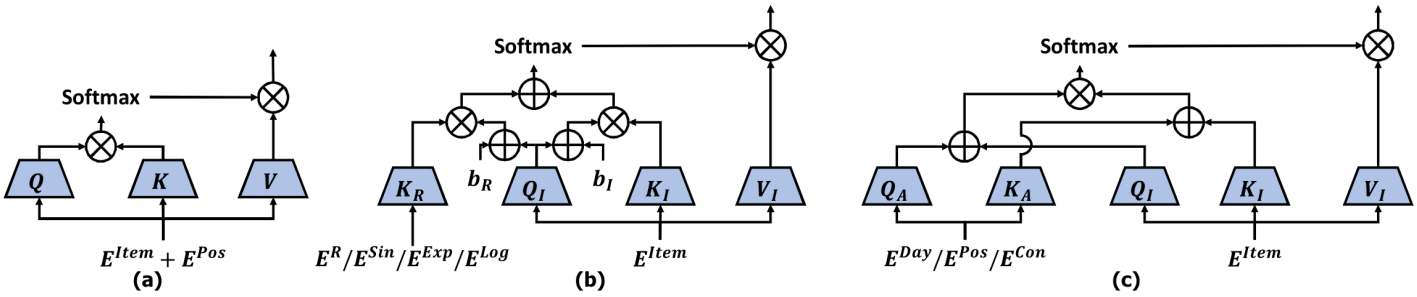

Fig. 2. Brief overview of single attention head in multi-headed attention structure. (a) absolute self-attention in Transformer (b)relative self-attention in Transformer-XL [2] and MEANTIME (c) absolute self-attention in MEANTIME.

E^{Pos} is replaced by the relative positional embedding E^R that only depends on the difference between the positional indices of the array, and positional information is no longer used at key and value encoder. Parameters b_I and b_R serve as global content/positional bias respectively

Quick Results

Table 1. Performance Comparison.

Datasets	Metrics	MARANK	SAS	TISAS	BERT	MEANTIME	Improvement
ML-1M	Recall@5	0.4427	0.5708	0.5869	0.5862	**0.6415**	9.31%
	Recall@10	0.5993	0.6966	0.7102	0.6987	**0.7412**	4.36%
	NDCG@5	0.3042	0.4142	0.4307	0.4420	**0.4932**	11.58%
	NDCG@10	0.3548	0.4550	0.4708	0.4786	**0.5255**	9.81%
ML-20M	Recall@5	0.3938	0.5274	0.5412	0.5910	**0.6225**	5.34%
	Recall@10	0.5503	0.6887	0.7021	0.7176	**0.7491**	4.39%
	NDCG@5	0.2665	0.3700	0.3814	0.4469	**0.4739**	6.04%
	NDCG@10	0.3170	0.4223	0.4336	0.4879	**0.5150**	5.55%
Beauty	Recall@5	0.1577	0.2011	0.2081	0.2271	**0.2470**	8.78%
	Recall@10	0.2302	0.2714	0.2805	0.3095	**0.3330**	7.58%
	NDCG@5	0.1085	0.1464	0.1512	0.1633	**0.1764**	8.05%
	NDCG@10	0.1318	0.1689	0.1745	0.1898	**0.2041**	7.52%
Game	Recall@5	0.2846	0.3925	0.3857	0.4211	**0.4752**	12.85%
	Recall@10	0.4217	0.5201	0.5085	0.5590	**0.6100**	9.12%
	NDCG@5	0.1893	0.2735	0.2735	0.2970	**0.3446**	16.03%
	NDCG@10	0.2334	0.3147	0.3133	0.3415	**0.3882**	13.67%

- All models are fully tuned through an extensive grid-search on hyperparameters: hidden dimension h ∈ {16,32,64,128}, number of heads n ∈ {2,4}, dropout ratio ∈ {0.2,0.5}, and weight decay wd ∈ {0,0.00001}.
- the best combination of embedding types among all possible combinations with n = 2,4: -
    - Day+Pos+Sin+Log for MovieLens 1M and 20M
    - Day+Sin+Exp+Log for Amazon Beauty
    - Day+Pos+Sin+Exp for Amazon Game.

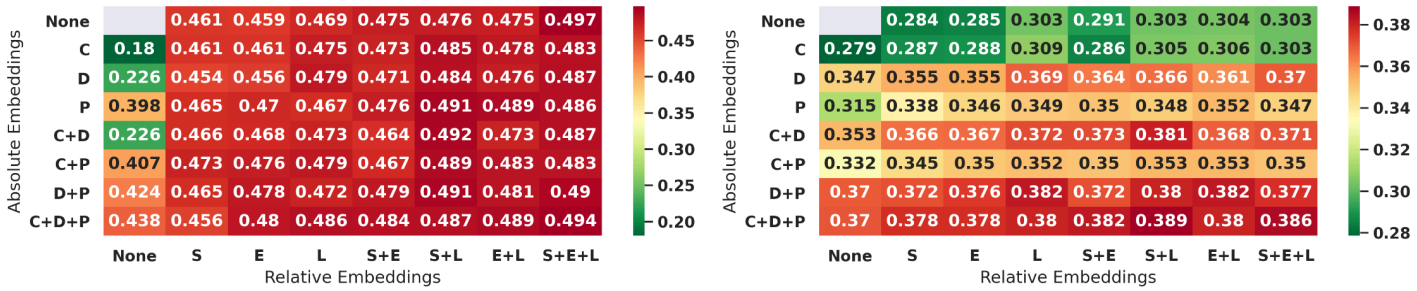

Fig. 3. NDCG@10 Comparison with All Possible Combinations of Positional Embeddings on ML-1M (left) and Game (right) Dataset.C, D, P, S, E and L represent Con ,Day, Pos, Sin, Exp and Log, repectively. We used h = 60 for all experiments.

Questions about the paper?

-

What do you like?

- They utilize fully timestamp information and use relative position concept with modified multi-head attention.
- It is interesting to use Sin/Log/Exp for representing various time patterns.
- They release well-structured source code with other baselines.

What you don't like?

-

How to improve / Any new ideas?

- Redefine the Relative distance matrix.
- Devise Item embedding method with timestamp

References

- [1] Li, Jiacheng, Yujie Wang, and Julian McAuley. 2020. “Time Interval Aware Self-Attention for Sequential Recommendation.” In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 322–30. WSDM ’20. New York, NY, USA: Association for Computing Machinery.
- [2] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive LanguageModels beyond a Fixed-Length Context. Association for Computational Linguistics (ACL)(2019).