# Self-Modulating Attention in Continuous Time Space with Applications to Sequential Recommendation

Chao Chen, Haoyu Geng,
Nianzu Yang, Junchi Yan,
Daiyue Xue, Jianping Yu, Xiaokang Yang

ICML
International Conference
On Machine Learning

## Abstract

User interests are usually dynamic in the real world, which poses both theoretical and practical challenges for learning accurate preferences from rich behavior data. Among existing user behavior modeling solutions, attention networks are widely adopted for its effectiveness and relative simplicity. Despite being extensively studied, existing attentions still suffer from two limitations: i) conventional attentions mainly take into account the spatial correlation between user behaviors, regardless the distance between those behaviors in the continuous time space; and ii) these attentions mostly provide a dense and undistinguished distribution over all past behaviors then attentively encode them into the output latent representations. This is however not suitable in practical scenarios where a user's future actions are relevant to a small subset of her/his historical behaviors. In this paper, we propose a novel attention network, named self-modulating attention, that models the complex and non-linearly evolving dynamic user preferences. We empirically demonstrate the effectiveness of our method on top-N sequential recommendation tasks, and the results on three large-scale real-world datasets show that our model can achieve state-of-the-art performance.

## Self-modulating Attention

The motivation for designing attention in continuous time space is to directly inform the correlation $p(v_{ij} \mid H_t)$ with the expected number of occurrence $E[N(t,t+dt) \mid \mid H_{t^-}, v_{ij}]$, conditional on the history $H_{t^-}$, where $H_{t^-}=H_t \cup \{t_j+1/ \in (t_j, t)\}$ and $N(t, t+dt) \in \{0,1\}$ denotes the number of occurrences for item $i_j$ in an infinitesimal interval.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{p(\mathbf{v}|\mathcal{H}_t)} \left[\mathbb{E}\left[\mathrm{N}(t, t+\mathrm{d}t) \mid \mathcal{H}_{t-}, \mathbf{v}\right] \mathbf{v}\right] \\
&\overset{(a)}{=} \mathbb{E}_{p(\mathbf{v}|\mathcal{H}_t)} \left[p\left(t_{j+1} \in [t, t+\mathrm{d}t] \mid \mathcal{H}_{t-}, \mathbf{v}\right) \mathbf{v}\right] \\
&\overset{(b)}{=} \mathbb{E}_{p(\mathbf{v}|\mathcal{H}_t)} \frac{p\left(t_{j+1} \in [t, t+\mathrm{d}t] \mid \mathcal{H}_t, \mathbf{v}\right)}{p\left(t_{j+1} \notin (t_j, t) \mid \mathcal{H}_t, \mathbf{v}\right)} \mathbf{v} \\
&\overset{(c)}{=} \mathbb{E}_{p(\mathbf{v}|\mathcal{H}_t)} \frac{f\left(t \mid \mathcal{H}_t, \mathbf{v}\right) \mathrm{d}t}{S\left(t \mid \mathcal{H}_t, \mathbf{v}\right)} \mathbf{v} \\
&\overset{(d)}{=} \mathbb{E}_{p(\mathbf{v}|\mathcal{H}_t)} \mathbf{v} \lambda^*\left(t \mid \mathcal{H}_t, \mathbf{v}\right) \mathrm{d}t
\end{aligned}
$$

The dynamic processes will be characterized by conditional intensity function $\lambda^*(t \mid H_t)$. The generalized self-modulating attention could be formulated as:

$$\widehat{\mathrm{Att}}\left(\mathbf{q}, \mid \mathcal{H}_t\right) = \mathbb{E}_{p\left(\mathbf{v}|_{\mathcal{H}_t}\right)} \mathbf{v} \lambda^*\left(t \mid \mathcal{H}_t, \mathbf{v}\right)$$

## Self-modulating Layer (SMLayer)

We introduce a specific implementation, self-modulating layer (SMLayer) with application to sequential recommendation.
The widely used self-attention in Sequential recommendation has following form:

$$
\begin{gathered}
\mathbf{X} = \mathrm{concat}([\mathbf{Y}, \mathbf{Z}]) \\
\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V}^{\mathrm{seq}} = \mathbf{X}\mathbf{W}^V \\
\mathbf{H} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}^{\mathrm{seq}}.
\end{gathered}
$$

where all W are projection matrices, Q, K, $V_{seq}$ are separately the query, key and values matrices obtained by different transformations of the input X, and H is the output representations of conventional attentions. Further we present the formulation of conditional intensity function:

$$\mathbf{g}_k(t) = \sigma(\underbrace{\mathbf{W}_k^G \mathbf{h}\left(t_j\right)}_{\text{Endogenous}} + \underbrace{\mathbf{b}_k^G\left(t - t_j\right)}_{\text{Exogenous}})$$

$$\lambda\left(t \mid \mathcal{H}_t, \mathbf{v}_k\right) = f_k\left(\mathbf{w}_k^\top \mathbf{g}_k(t) + \mu_k\right)$$

where $g_k(t)$ signifies continuous time temporal dynamics, and $f_k$ is the softplus activation function:

$$f_k(x) = \phi_k \log\left(1 + \exp\left(x/\phi_k\right)\right)$$

## Continuous Time Regularization (CTReg)

In preference learning, the model parameters are usually optimized by minimizing the reconstruction error. The only source of supervised signal is from the behavior data (i.e.,R) that is independent of time Intensity function learned in protocol above might probably diverge from the complex continuous-time patterns contained in the data. CTReg (continuous time regularization) is defined by:

$$R(\Theta) = \sum_{j=1}^{L} \log \lambda_k\left(t_j \mid \mathcal{H}_j\right) - \int_{t_1}^{t_L} \lambda\left(t_j \mid \mathcal{H}_j\right) \mathrm{d}t$$

The final objective function is minimization of empirical risk with CTReg as:

$$\min_{\Theta} \ell(\mathbf{R}, \hat{\mathbf{R}}) - \gamma \mathbb{E}_{u \in [1,m]} R(\Theta; u)$$

## Theoretical Results

Generalization Bound. Suppose that the loss function is L-Lipschitz, and for the estimate R on an random example set Ω, we bound $\rho|\Omega| = \sup_{P \in F} \sum_{(u,i) \in \Omega} \|P_{i,*}\|_0$ and $\mu = \sup_{(u,i,k) \in \Omega} |P_{u,k}(VB)_{k,j}|$, then with probability at least 1−δ, we have the bound:

$$\mathbf{E}[l(\mathbf{R}, \hat{\mathbf{R}})] \leq \mathbf{E}_\Omega[l(\mathbf{R}, \hat{\mathbf{R}})] + \mathcal{O}\left(L\mu\sqrt{\frac{C_p \ln|\Omega|}{|\Omega|}} + \sqrt{\frac{\ln(1/\delta)}{|\Omega|}}\right)$$

$$\text{where } C = d(m+n)\log(48emn)$$

## Empirical Results

### Quantitative Analysis

We present experimental results on three large-scale datasets (Amazon, Koubei, Tmall) against sequential SOTA recommendation baselines.

| Model | Amazon | | Koubei | | Tmall | |
|---|---|---|---|---|---|---|
| | HR | NDCG | HR | NDCG | HR | NDCG |
| **SHAN** | 0.19250 | 0.11724 | 0.28150 | 0.20256 | 0.37316 | 0.25840 |
| **DIN** | 0.21955 | 0.13443 | 0.32665 | 0.24186 | 0.48460 | 0.33855 |
| **GRU4REC** | 0.24380 | 0.15822 | 0.32655 | 0.27052 | 0.46877 | 0.33746 |
| **SASREC** | 0.25595 | 0.16131 | 0.35455 | 0.27070 | 0.50433 | 0.34326 |
| **SASREC+** | 0.25820 | 0.16204 | 0.35690 | 0.27148 | 0.50607 | 0.34328 |
| **SASREC w/ ours.** | **0.26545** | **0.16529** | **0.36235** | **0.28083** | **0.51347** | **0.35811** |

Comparison to state-of-the-art Baselines on three benchmark Datasets

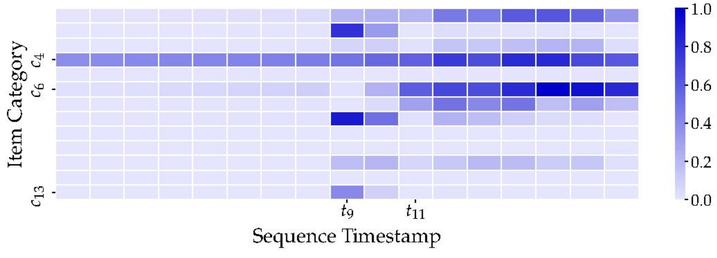
We further conduct ablation study to equip generalized self-modulating attention on DIN (Deep interest network) with proposed SMLayer and CTReg (Continuous time regularization) to validate our methods:

| | Dataset | DIN | | | SASREC | | |
|---|---|---|---|---|---|---|---|
| | | Origin | +SMLayer | +CTReg | Origin | +SMLayer | +CTReg |
| HR | **Amazon** | 0.21955 | **0.22065** | 0.21985 | 0.25595 | **0.26545** | 0.26058 |
| | **Koubei** | 0.32665 | **0.33940** | 0.33780 | 0.35455 | 0.36194 | **0.36235** |
| | **Tmall** | 0.48460 | 0.49033 | **0.49157** | 0.50433 | 0.51218 | **0.51347** |
| NDCG | **Amazon** | 0.13443 | **0.13383** | 0.13296 | 0.16131 | 0.16475 | **0.16529** |
| | **Koubei** | 0.24186 | **0.25444** | 0.25411 | 0.27070 | 0.27862 | **0.28083** |
| | **Tmall** | 0.33855 | 0.34580 | **0.35062** | 0.34326 | 0.35214 | **0.35811** |

Ablation Study

### Qualitative Analysis

Example of user preference intensities on Koubei. The user usually purchases the items from category c4 for all the time, and he/she starts to repeatedly buy the items from category c6 after time t9. The darker color corresponds to the higher intensity




### References

Kang, W.-C. and McAuley, J. Self-attentive sequential recommendation. In Proceedings of the IEEE International Conference on Data Mining (ICDM '18), pp. 197–206.IEEE, 2018

Xu, D., Ruan, C., Kumar, S., Korpeoglu, E., and Achan, K. Self-attention with functional time representation learning. In Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS '19),pp. 15915–15925, 2019

Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., Yan, Y., Jin, J., Li, H., and Gai, K. Deep interest network for click-through rate prediction. In Proceedings of the

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '18), pp. 1059–1068, 2018