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

Chao Chen, Haoyu Geng, Nianzu Yang, Junchi Yan,

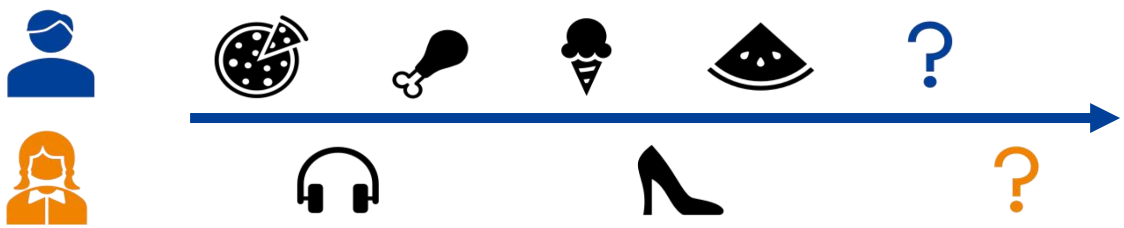
Daiyue Xue, Jianping Yu, Xiaokang Yang

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Background and Motivation

- **Challenge of attentions in Continuous Time Space:**
  - Attention models sequential positions, regardless of continuous timestamps
  - Attention provides dense distribution over behaviors, different from its sparse nature

- **Our contribution**:
  - Generalize regular attention to continuous time space
  - Propose self-modulating layer (SMLayer) to model spatial-temporal dynamics
  - Propose continuous time regularization (CTReg) to fit time-dependent patterns



User evolving behavior modeling

Generalize Attention to Continuous Time Space

- **General attention**

$$\mathrm{Att}\left(\mathbf{q} \mid \mathcal{H}_j\right) = \sum_{\mathbf{v}} p\left(\mathbf{v} \mid \mathcal{H}_j\right) \mathbf{v} = \mathbf{E}_{p\left(\mathbf{v} \mid \mathcal{H}_j\right)} \mathbf{v}$$

- **Self-Modulated attention**

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

where λ^* is conditional intensity function of temporal point process

- **Properties of self-modulating attention**
  I. When $\lambda^*(t \mid H_t, v_i)=0$, the impact of event v_i is rejected.
  II. When $\lambda^*(t \mid H_t, v_i)<1$, the impact of event v_i is attenuated.
  III. When $\lambda^*(t \mid H_t, v_i)>1$, the impact of event v_i is amplified.

Where:
Ht: historical interactions
N(tj, tj +dt): the number of occurrences for item i_j in an infinitesimal interval
F: *D*ensity function
F (*t* |*Ht;* **v**): Cumulative distribution function
S(*t* |*Ht;* **v**): Survival function = 1-F(*t* |*Ht;* **v**)
*λ**(t I Ht*;* **v**): conditional intensity function

$$\lambda^*\left(t \mid \mathcal{H}_t, \mathbf{v}\right) = \frac{f\left(t \mid \mathcal{H}_t, \mathbf{v}\right)}{1 - F\left(t \mid \mathcal{H}_t, \mathbf{v}\right)}$$

Self-modulating Layer (SMLayer)

- **Impact of sequential positions**
  - Y is historical embedding， Z is positional encoding
  - Q, K, V^{seq} are query, key, value, and H is self attention outputs

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

- **Impact of temporal timestamps**
  - Endogenous: influence within the sequence
  - Exogenous: forces that reacts on the potential next item k during the time interval
  - Build conditional intensity function
    - f_k is softplus activation function

$$\mathbf{g}_k(t) = \sigma(\underbrace{\mathbf{W}_k^G \mathbf{h}(t_j)}_{\text{Endogenous}} + \underbrace{\mathbf{b}_k^G (t - t_j)}_{\text{Exogenous}})$$
$$\lambda(t \mid \mathcal{H}_t, \mathbf{v}_k) = f_k\left(\mathbf{w}_k^\top \mathbf{g}_k(t) + \mu_k\right)$$
$$f_k(x) = \phi_k \log(1 + \exp(x/\phi_k))$$

- **Combination of positions and timestamps**

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

Continuous Time Regularization

- **Continuous time regularization** (CTReg)
  - The only supervision signal comes from user historical behaviors, irrelevant to time interval

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

where λ_k is conditional intensity for type k, and λ is log-survival probabilities:

$$\lambda\left(t_j \mid \mathcal{H}_{t_j}^{(u)}\right) = \sum_k \lambda_k\left(t_j \mid \mathcal{H}_j^{(u)}\right)$$

- **Overall objective function:**

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

Experiments

- **Sequential Recommendations**
  - Datasets: Amazon, Koubei, Tmall
  - Strong generalization Protocol
  - Evaluation metrics:

    Hit Rate(HR), Normalized Discounted Cumulative Gain (NDCG)

Dataset	#Users	#Items	#Interactions
Amazon	211,384	18,490	1.6M
Koubei	212,831	10,213	1.8M
Tmall	320,497	21,876	7,6M

- **Ablation Studies**

		DIN			SASREC		
	Dataset	Origin	+SMLayer	+CTReg	Origin	+SMLayer	+CTReg
HR	**Amazon**	0.21955	**0.22065**	0.21985	0.25595	**0.26545**	0.26058
HR	**Koubei**	0.32665	**0.33940**	0.33780	0.35455	0.36194	**0.36235**
HR	**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**
NDCG	**Koubei**	0.24186	**0.25444**	0.25411	0.27070	0.27862	**0.28083**
NDCG	**Tmall**	0.33855	0.34580	**0.35062**	0.34326	0.35214	**0.35811**

Ablation study on the Amazon, Koubei and Tmall datasets. The proposed self-modulating layer (SMLayer) and continuous-time regularization (CTReg) are adapted to attention-based and SASREC models. The performance is evaluated in terms of HR@10 and NDCG@10.

Experiments

▪ Performance Comparison

SOTA baselines: SHAN [1], DIN[2], GRU4REC[3], SASREC[4], SASREC+[5]

	Amazon		Koubei		Tmall	
Model	HR	NDCG	HR	NDCG	HR	NDCG
SHAN (Ying et al., 2018)	0.19250	0.11724	0.28150	0.20256	0.37316	0.25840
DIN (Zhou et al., 2018)	0.21955	0.13443	0.32665	0.24186	0.48460	0.33855
GRU4REC (Hidasi et al., 2016)	0.24380	0.15822	0.32655	0.27052	0.46877	0.33746
SASREC (Kang & McAuley, 2018)	0.25595	0.16131	0.35455	0.27070	0.50433	0.34326
SASREC+ (Xu et al., 2019)	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**

Performance comparison between the baselines and our proposed method on the Amazon, Koubei and Tmall datasets in terms of HR@10 and NDCG@10. Boldfaces mean that the method performs statistically significantly better under t-tests, at the level of 95% confidence level. We emphasize the comparison against SASREC+, a variant of SASREC equipped with functional time embedding which captures continuous-time temporal dynamics.

[1] Ying, H., Zhuang, F., Zhang, F., Liu, Y., Xu, G., Xie, X.,Xiong, H., and Wu, J. Sequential recommender system based on hierarchical Attention networks. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '18), 2018
[2] 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. 1059–068, 2018.
[3] Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D.Session-based recommendations with recurrent neural networks. In Proceedings of the International Conference on Learning Representations (ICLR '16), 2016
[4] 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
[5] 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.

Thanks!