SELF-ATTENTION WITH FUNCTIONAL TIME REPRESENTATION LEARNING

Da Xu*, Chuanwei Ruan*, Sushant Kumar, Evren Korpeoglu, Kannan Achan

Walmart Labs

Self-attention for Continuous-time Sequence Modelling?

- Self-attention mechanism [4] is powerful but only works on discrete-time sequence with positional encoding.
- Time spans between sequential events often carry important signals.
- We identify the forms of functional time mapping that work well with self-attention especially the scaled dot-product attention.
- The proposed approaches have solid theoretical justification and guarantees. Experiments demonstrate its great practical values on real-world continuous-time sequence datasets.

Preliminaries in Functional Analysis

Temporal kernel. Embedding time from an interval $T = [0, t_{\max}]$ to $\mathbb{R}^d$ is equivalent to finding a mapping $\Phi : T \to \mathbb{R}^d$. Due to the **inner product** formulation of self-attention [4] and the **translation invariant** property of time spans, we define the *temporal kernel* as $\mathcal{K} : T \times T \to \mathbb{R}$ where $\mathcal{K}(t_1, t_2) := \langle \Phi(t_1), \Phi(t_2) \rangle$ and $\mathcal{K}(t_1, t_2) = \psi(t_1 - t_2), \forall t_1, t_2 \in T$ for some $\psi : [-t_{\max}, t_{\max}] \to \mathbb{R}$.

Embedding as feature maps. The feature map Φ captures how the temporal kernel function embeds the original time data into a higher dimensional space. So the task of learning temporal patterns is converted to a kernel learning problem with Φ as feature map.

Theorem 1 (**Bochner's Theorem** [1]). *A continuous, translation-invariant kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x} - \mathbf{y})$ on $\mathbb{R}^d$ is positive definite if and only if there exists a non-negative measure on $\mathbb{R}$ such that ψ is the Fourier transform of the measure.*

Implications: when scaled properly we can express $\mathcal{K}$ with:

$$\mathcal{K}(t_1, t_2) = \psi(t_1, t_2) = \int_{\mathbb{R}} e^{i\omega(t_1 - t_2)} p(\omega) d\omega = E_\omega\big[\xi_\omega(t_1)\xi_\omega(t_2)^*\big], \tag{1}$$

where $\xi_\omega(t) = e^{i\omega t}$. Since the kernel $\mathcal{K}$ and the probability measure $p(\omega)$ are real, we extract the real part of (1) and obtain an alternate expression of the kernel:

$$\mathcal{K}(t_1, t_2) = E_\omega\big[\cos(\omega(t_1 - t_2))\big] = E_\omega\big[\cos(\omega t_1)\cos(\omega t_2) + \sin(\omega t_1)\sin(\omega t_2)\big]. \tag{2}$$

Theorem 2 (**Mercer Theorem** [2]). *Consider the function class $L^2(\mathcal{X}, \mathbb{P})$ where $\mathcal{X}$ is compact. Suppose that the kernel function $\mathcal{K}$ is continuous with positive semidefinite and satisfy the condition $\int_{\mathcal{X}\times\mathcal{X}} \mathcal{K}^2(x, z) d\mathbb{P}(x) d\mathbb{P}(y) \leq \infty$, then there exist a sequence of eigenfunctions $(\phi_i)_{i=1}^\infty$ that form an orthonormal basis of $L^2(\mathcal{X}, \mathbb{P})$, and an associated set of non-negative eigenvalues $(c_i)_{i=1}^\infty$ such that*

$$\mathcal{K}(x, z) = \sum_{i=1}^{\infty} c_i \phi_i(x) \phi_i(z), \tag{3}$$

where the convergence of the infinite series holds absolutely and uniformly.

Implications: we can embed instances from the functional time domain T into the infinite sequence space $\ell^2(\mathbb{N})$, by defining the mapping via $t \mapsto \Phi^{\mathcal{M}}(t) := \big[\sqrt{c_1}\phi_1(t), \sqrt{c_2}\phi_2(t), \dots\big]$, and Mercer's Theorem guarantees the convergence of $\big\langle \Phi^{\mathcal{M}}(t_1), \Phi^{\mathcal{M}}(t_2) \big\rangle \to \mathcal{K}(t_1, t_2)$.

Note. We still haven't reached a feasible parametric form for Φ, since the $p(\omega)$ in Bochner's and the set of basis $\{\phi_i\}$ in Mercer's are unknown.

Proposed Approaches

Bochner's encoding. Following the implications of *Bochner's* Theorem, the expectation in (1) can be approximated by Monte Carlo integral [3]. With d samples drawn from $p(\omega)$, an estimate of our kernel $\mathcal{K}(t_1, t_2)$ can be constructed by $\frac{1}{d}\sum_{i=1}^{d} \cos(\omega_i t_1)\cos(\omega_i t_2) + \sin(\omega_i t_1)\sin(\omega_i t_2)$. So we propose finite dimensional Bochner feature map:

$$t \mapsto \Phi_d^{\mathcal{B}}(t) := \sqrt{\frac{1}{d}}\big[\cos(\omega_1 t), \sin(\omega_1 t), \dots, \cos(\omega_d t), \sin(\omega_d t)\big],$$

and we prove the following claim which guarantees the stochastic uniform convergence.

Claim 1. *Let $p(\omega)$ be the corresponding probability measure stated in Bochner's Theorem for kernel function $\mathcal{K}$. Suppose the feature map Φ is constructed as described above using samples $\{\omega_i\}_{i=1}^d$, we have*

$$Pr\Big(\sup_{t_1, t_2 \in T} \big|{\Phi_d^{\mathcal{B}}(t_1)}' \Phi_d^{\mathcal{B}}(t_2) - \mathcal{K}(t_1, t_2)\big| \geq \epsilon\Big) \leq 4\sigma_p \sqrt{\frac{t_{\max}}{\epsilon}} exp\Big(\frac{-d\epsilon^2}{32}\Big), \tag{4}$$

where σ_p^2 is the second momentum with respect to $p(\omega)$.

Therefore, we can either use parametric or non-parametric distributional learning methods to obtain samples from the optimized $p(\omega)$, and then construct $\Phi_d^{\mathcal{B}}$ accordingly.

Mercer's encoding. As for the *Mercer* Theorem, we prove in the following Proposition 1 that a straightforward parameterization of the feature map via the Fourier basis expansion is possible, by decomposing the temporal kernel $\mathcal{K}$ into a set of periodic kernel functions $\{\mathcal{K}_\omega\}$.

Proposition 1. *For kernel function $\mathcal{K}$ that is continuous, PSD and translation-invariant with $\mathcal{K} = \psi(t_1 - t_2)$, suppose ψ is a even periodic function with frequency ω, i.e $\psi(t) = \psi(-t)$ and $\psi\big(t + \frac{2k}{\omega}\big) = \psi(t)$ for all $t \in [-\frac{1}{\omega}, \frac{1}{\omega}]$ and integers $k \in \mathbb{Z}$, the eigenfunctions of $\mathcal{K}$ are given by the Fourier basis.*

After truncating the series of Fourier basis, we have the infinite dimensional Mercer's feature map for each $\mathcal{K}_\omega$:

$$t \mapsto \Phi_\omega^{\mathcal{M}}(t) = \Big[\sqrt{c_1}, \dots, \sqrt{c_{2j}}\cos\big(\frac{j\pi t}{\omega}\big), \sqrt{c_{2j+1}}\sin\big(\frac{j\pi t}{\omega}\big), \dots\Big],$$

where c_j are the corresponding Fourier coefficients and ω_i are free model parameters.

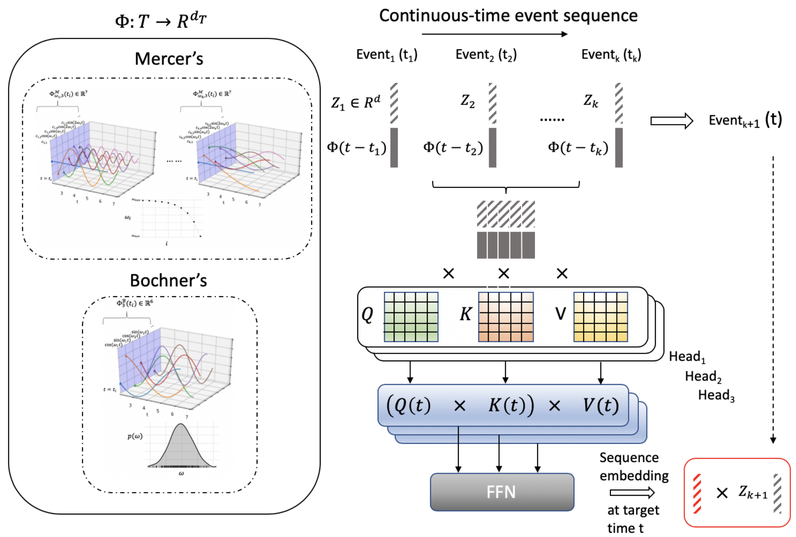

Fig. 1: Left panel: visual illustration of the proposed Bochner and Mercer time embedding ($\Phi_d^{\mathcal{B}}(t)$ and $\Phi_{\omega,d}^{\mathcal{M}}(t)$) for a specific $t = t_i$ with $d = 3$. Right panel: network architecture for next-event prediction at time t with a single block.

Experiments and Results

Time-event interaction: we first concatenate the event embedding and time representations into $[\mathbf{Z}, \mathbf{Z}_T]$ where $\mathbf{Z} = \big[Z_1, \dots, Z_q\big]$, $\mathbf{Z}_T = \big[\Phi(t_1), \dots, \Phi(t_q)\big]$ and then project them into the $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ spaces respectively to capture their linear or non-linear interactions, e.g.

$$\mathbf{Q} = \text{ReLU}\big([\mathbf{Z}, \mathbf{Z}_T]\mathbf{W}_0 + b_0\big)\mathbf{W}_1 + b_1,$$

and finally we use $\mathbf{h}^{(i)} = \text{Attn}^{(i)}\big(\mathbf{Q}, \mathbf{K}, \mathbf{V}\big)$ as the hidden output of the i^{th} head in the multi-head attention setting.

Feature maps specified by $[\phi_{2i}(t), \phi_{2i+1}(t)]$	Origin	Parameters	Interpretations of ω
$\big[\cos\big(\omega_i(\mu)t\big), \sin\big(\omega_i(\mu)t\big)\big]$	Bochner Normal	μ: location-scale parameters specified for the *reparametrization trick*.	$\omega_i(\mu)$: converts the i^{th} sample (drawn from auxiliary distribution) to target distribution under location-scale parameter μ.
$\big[\cos\big(g_\theta(\omega_i)t\big), \sin\big(g_\theta(\omega_i)t\big)\big]$	Bochner Inv CDF	θ: parameters for the inverse CDF $F = g_\theta$.	ω_i: the i^{th} sample drawn from the auxiliary distribution.
$[\cos(\bar{\omega}_i t), \sin(\bar{\omega}_i t)]$	Bochner Non-param	$\{\bar{\omega}\}_{i=1}^d$: transformed samples under non-parametric inverse CDF transformation.	$\bar{\omega}_i$: the i^{th} sample of the underlying distribution $p(\omega)$ in Bochner's Theorem.
$\big[\sqrt{c_{2i,k}}\cos(\omega_j t), \sqrt{c_{2i+1,k}}\sin(\omega_j t)\big]$	Mercer	$\{c_{i,k}\}_{i=1}^{2d}$: the Fourier coefficients of corresponding $\mathcal{K}_{\omega_j}$, for $j = 1, \dots, k$.	ω_j: the frequency for kernel function $\mathcal{K}_{\omega_j}$ (can be parameters).

Fig. 2: A summary of the proposed approaches.

We compare among the proposed function mapping methods (the details are shown in Figure 2) in self-attention and with state-of-the-art baseline approaches including self-attention with positional encoding (PosEnc), on recommendation tasks with the **Stack Overflow**, **Moivelens** and **Walmart.com** datasets. Case studies and analysis on the attention weights are given in Fig. 3. The results are provided in Fig. 4.

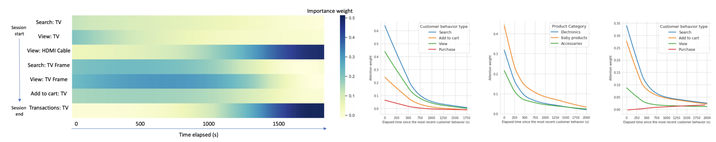

Fig. 3: Attention weight analysis as functions of time and event.

Stack Overflow								
Method	LSTM	TimeJoint	RMTPP	PosEnc	**Bochner Normal**	**Bochner Inv CDF**	**Bochner Non-para**	**Mercer**
Accuracy	46.03(.21)	46.30(.23)	46.23(.24)	44.03(.33)	44.89(.46)	44.67(.38)	46.27(0.29)	**46.83(0.20)**
config						$\overline{NVP}$		$k = 10$
MovieLens-1m								
Method	GRU4Rec	Caser	TransRec	-	-	-	-	-
Hit@10	75.01(.25)	78.86(.22)	64.15(.27)	82.45(.31)	81.60(.69)	82.52(.36)	82.86(.22)	**82.92 (.17)**
NDCG@10	55.13(.14)	55.38(.15)	39.72(.16)	59.05(.14)	59.47(.56)	60.80(.47)	60.83(.15)	**60.88 (.11)**
config						$\overline{MAF}$		$k = 5$
Walmart.com data								
Method	GRU4Rec	RNN+attn	TransRec	-	-	-	-	-
Hit@5	4.12(.19)	5.90(.17)	7.03(.15)	8.63(.16)	4.27(.91)	9.04(.31)	9.25(.15)	**10.92(.13)**
NDCG@5	4.03(.20)	4.66(.17)	5.62(.17)	6.92(.14)	4.06(.94)	7.27(.26)	7.34(.12)	**8.90(.11)**
Hit@10	6.71(.50)	9.03(.44)	10.38(.41)	12.49(.38)	7.66(.92)	12.77(.65)	13.16(.41)	**14.94(.31)**
NDCG@10	4.97(.31)	7.36(.26)	8.72(.26)	10.84(.26)	6.02(.99)	10.95(.74)	11.36(.27)	**12.81(.22)**
config						$\overline{MAF}$		$k = 25$

Fig. 4: Performance metrics for the proposed approach and baseline models. *MAF* and *NVP* are the flow-based distribution learning methods, and k gives the dimension of Fourier basis expansions.

[1] Lynn H Loomis. *Introduction to abstract harmonic analysis*. Courier Corporation, 2013.

[2] James Mercer. "Xvi. functions of positive and negative type, and their connection the theory of integral equations". In: *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character* 209.441-458 (1909), pp. 415–446.

[3] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems*. 2008, pp. 1177–1184.

[4] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.