Koopman Attention: Dynamical system trajectories of query updates

Preprint · February 2024

DOI: 10.13140/RG.2.2.12221.64486

1 author:

Gary Nan Tie

Mu Risk LLC

Koopman Attention:
Dynamical system trajectories of query updates

Gary Nan Tie
Feb 26, 2024

Abstract
A query-key-value attention mechanism, where query updates are performed by perturbed Koopman operator state predictions, instead of feed-forward network layers in Transformers, introduces stochastic nonlinearity to query update trajectories. This gives a new foundation for large language models.

Keywords: attention, large language models, Koopman operator, nonlinear dynamical system, Representer Theorem

Attention mechanism for sequence-to-sequence or next-token:

Represent each word by a unique token, a real valued N-vector.

Project tokens (Tok) into a lower dimensional space,

the so called latent space (Lat) of n-vectors, $n < N$,

using some learnt weight matrices W to be specified.

Given source tokens $s_1, \ldots, s_S$ and

target tokens $t_1, \ldots, t_T$

For Encoder-Decoder cross attention; for self-attention

source and target tokens are the same.

Define latent vectors:

initial query $q_i^0 = W_Q t_i$, $i \in [T]$

key $k_j = W_K s_j$, $j \in [S]$

value $v_j = W_V s_j$, $j \in [S]$

where

$$W_Q, W_K, W_V : \overset{\mathbb{R}^N}{\overset{\cup}{\mathrm{Tok}}} \longrightarrow \overset{\mathbb{R}^n}{\overset{\cup}{\mathrm{Lat}}}$$

For each position $i \in [T]$,

let $q_i^0, q_i^1, \ldots, q_i^{m-1} \in \mathcal{X} \subseteq \mathbb{R}^n$ be a trajectory

of query updates. We now describe how to obtain the

next update $q_i^m \in \mathbb{R}^n$. (For Encoder self-attention;

for Decoder also apply future masking).

Let $a_{ij} = \text{similarity}(q_i^{m-1}, k_j)$, $i \in [T]$, $j \in [S]$,

(In self-attention, each token can attend all other tokens,

enabling long range dependencies.)

where similarity is a scaled dot product or more generally

a reproducing Kernel and let probability $\alpha_{ij} = \text{softmax}(a_{ij})$.

Define $q_i' = \sum_{j=1}^{S} \alpha_{ij} v_j \in \mathbb{R}^n$, $i \in [T]$

(Notice the target information persists through the α's.)

and let s_{ij*} denote the s_j with the highest probability α_{ij}.

Let $\mathbb{k}_2 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be any desired kernel,

then reproducing kernel $\mathbb{k}(x,y) \triangleq \sum_{p=1}^{n} x_p y_p + \mathbb{k}_2(x,y)$

satisfies $\begin{aligned} \pi_p : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ (z_1,\dots,z_n) &\mapsto z_p \end{aligned} \quad \in H(\mathbb{k}) \quad \forall p \in [n]$

and $H(\mathbb{k}_2) \subseteq H(\mathbb{k})$.

So let $\mathbb{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel

such that $\forall p \in [n]$, $\pi_p \in H(\mathbb{k})$.

For each key k_j, $j \in [S]$ define observable map

$g_j \in H(\mathbb{k}) \subseteq \mathrm{Fun}(\mathcal{X}, \mathbb{R})$ by $g_j = \mathbb{k}(k_j, \cdot)$

For each position $i \in [T]$, we have the following Koopman data [NanTie]:

a) query update trajectory $q_i^0, q_i^1, \dots, q_i^{m-1} \in \mathbb{R}^n$

b) observable maps $g_1, g_2, \dots, g_S \in H(\mathbb{k})$

where $g_j = \mathbb{k}(k_j, \cdot)$ for key k_j

c) measurements $y_{jl} \triangleq g_j(q_i^l) = \mathbb{k}(k_j, q_i^l)$

for $j = 1, \dots, S$ and $l = 0, \dots, m-1$.

Following [Nan Tie, 'Koopman transfer learning via perturbation']

① Determine vector-valued Koopman operator (that depends on the trajectory of prior query updates), $\mathbb{K} \triangleq \bigoplus_{p \in [n]} \hat{K}_p : \mathcal{W}^n \longrightarrow \mathcal{W}^n$ from the constrained Representer Theorems for $\hat{K}_p : \mathcal{W} \longrightarrow \mathcal{W}$.

② Determine Koopman embedding $\varphi \in \mathcal{W}^n$,

$$\varphi = [\varphi_p] : \mathcal{X} \longrightarrow \mathbb{R}^n \quad \text{where } \varphi_p \triangleq \hat{K}_p \pi_p \quad \text{for } p \in [n].$$

③ To proxy φ, draw* invertible perturbation $\psi \triangleq [\mathbb{K}_p \pi_p]$

④ Define m-th updated query, for position $i \in [T]$ by:

$$q_i^m \triangleq \psi^{-1}(\hat{\mathbb{K}} \varphi(q_i^1)) \in \mathbb{R}^n$$

(instead of using a feed-forward neural network).

* Perturbation ψ is a function of two hyperparameters $\lambda > 0$ and $\varepsilon > 0$, draw them from a distribution on $(0, 1]$ concentrated at 0. Hence the nonlinear query update trajectory $q_i^0, q_i^1, \ldots, q_i^m$ is stochastic.

Hence for sequence-to-sequence attention:

$t_1, \ldots, t_T$ generates after m query updates

$q_1^m, \ldots, q_T^m$ which predicts output tokens

$o_2, \ldots, o_{T+1}$ where $o_i = s_{(i-1)j^*}$ with probability $\alpha_{(i-1)j^*}$

(and for inference apply $\tilde{W}: Lat \to Tok$ to create

updated targets $t_i' \triangleq \tilde{W} q_i^m$)

For next-token inference:

$t_1 =$ empty token or seed like t_T,

$t_2 = s_{1j^*}$, $t_3 = s_{2j^*}$, $\ldots$ autoregressive updates.

Summary: Instead of feed-forward network layers used in Transformers, we introduce novel Koopman state prediction to update queries in attention mechanisms.

References

Nan Tie, G.
'Koopman Transfer Learning via Perturbation'
DOI: 10.13140/RG.2.2.21890.66248/3
https://rgdoi.net/10.13140/RG.2.2.21890.66248/3
Jan, 2024.

Nan Tie, G.
'Query Attention'
DOI: 10.13140/RG.2.2.17359.56488
https://www.researchgate.net/publication/370364025
Apr, 2023.

Khosravi, M.
'Representer Theorem for Learning Koopman Operators'
IEEE Transactions on Automatic Control,
Vol. 68, No. 5, May 2023

V. Paulsen, M. Raghupathi,
"An Introduction to the Theory of Reproducing Kernel Hilbert Spaces",
Cambridge University Press, 2016.
ISBN 978-1-107-10409-9