Choose a Transformer: Fourier or Galerkin

Shuhao Cao
Washington University in St. Louis

Overview on the shoulder of Giants

- Ever since the landmark [1], everyone is Transformer'ing, yet the mathematics behind the attention mechanism is not well-understood.
- Standing on the shoulder of Giants, the Fourier Neural Operator [2], the latent representation is interpreted "column-wise" (each column represents a basis on a discrete grid), opposed to the conventional "row-wise"/ "position-wise"/"word-wise" interpretation of the attention in Natural Language Processing (NLP).
- Softmax normalization, or its kernelized approximation, is not a necessary component in encoder-only models.
- The Galerkin-type attention (a linear attention without softmax) has an architectural approximation capacity to represent explicitly a Petrov-Galerkin projection under a Hilbertian setup.
- The Galerkin-type attention operator has a "translation" capacity to represent

$$\min_{\text{values}} \|\mathfrak{a}(\text{query},\cdot) - \mathfrak{b}(\cdot,\text{values})\|_{\text{dual space of keys}} = \min_{v} \sup_{k\in\text{keys}} \frac{|\mathfrak{a}(q,k)-\mathfrak{b}(k,v)|}{\|k\|},$$

thus to learn a latent representation space on which the input (query) and the output (values) are "close", and this closeness is measured by how they respond to dynamically changing keys (functional norm).

- Replacing half of the trainable parameters in FNO by Galerkin Transformer encoder, the model's evaluation accuracy is significantly improved in PDE solution operator learning benchmark problems. Galerkin Transformer encoder-only model is capable of recovering coefficients (inverse coefficient identification) based on noisy measurements that traditional methods or FNO cannot accomplish.

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is All you Need*, in: *Advances in Neural Information Processing Systems (NIPS 2017)*, vol. 30, 2017

[2] Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, *Fourier Neural Operator for Parametric Partial Differential Equations*, in: *International Conference on Learning Representations*, 2021, URL: https://openreview.net/forum?id=c8P9NQVtmnO

Attention is an Operator Learner

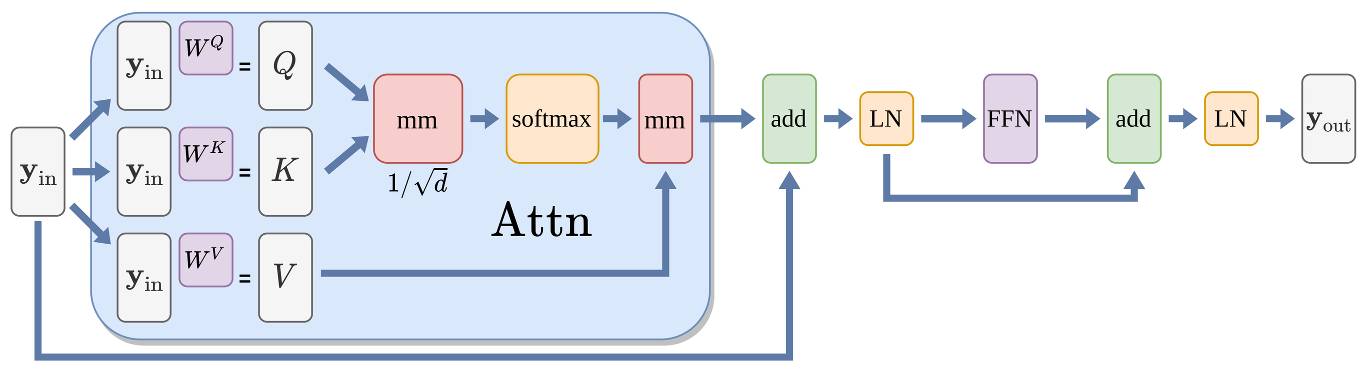


Fig. 1: Self-attention mechanism in the classical Transformer (figure reproduced from *Attention is All You Need*).

- $\mathbf{y}_{\text{in}} := \mathbf{y}, \mathbf{y}_{\text{out}} \in \mathbb{R}^{n\times d}$, input/output sequences; positional encodings added.
- Latent representations: query Q, key K, value V generated by 3 trainable matrices $W^Q, W^K, W^V \in \mathbb{R}^{d\times d}$: $Q = \mathbf{y}W^Q$, $K = \mathbf{y}W^K$, $V = \mathbf{y}W^V$.
- For $\text{Attn}_s(\mathbf{y}) := \text{Softmax}\big(d^{-1/2}QK^\top\big)V$. The full non-masked self-attention can learn a map with an input being a variable length latent representation and output being another same length latent representation

$$\text{Attn} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}, \quad \mathbf{z} = \mathbf{y} + \text{Attn}_s(\mathbf{y}), \quad \mathbf{y} \mapsto \text{Ln}\big(\mathbf{z} + g\big(\text{Ln}(\mathbf{z})\big)\big) =: \mathbf{y}_{\text{out}}.$$

- The softmax succeeding the matrix multiplication convexifies the weights for combining different positions to enable a kernel interpretation. However, softmax acts globally thus slow.
- Fourier Neural Operator [2] exerts FFT/iFFT (change of bases) for column bases of the latent representations, uses the natural $1/\sqrt{n}$ normalization: no softmax!

Rethinking the latent representation in Hilbert spaces

- Re-interpreting the latent representation in $\mathbb{R}^{n\times d}$ from:

  Row = A word to Column = A basis function in a Hilbert subspace.

- The columns of query/keys/values contain the learned basis functions spanning certain subspaces of different Hilbert spaces.
- The latent approximation spaces will be enriched by $\text{span}\{w_j \in \mathbb{X}_h : w_j(x_i) = (\sigma_s(\mathbf{x}))_{ij}, 1 \le j \le d\} \subset \mathscr{H}$ to try to capture how an operator of interest responses to the subset of inputs.

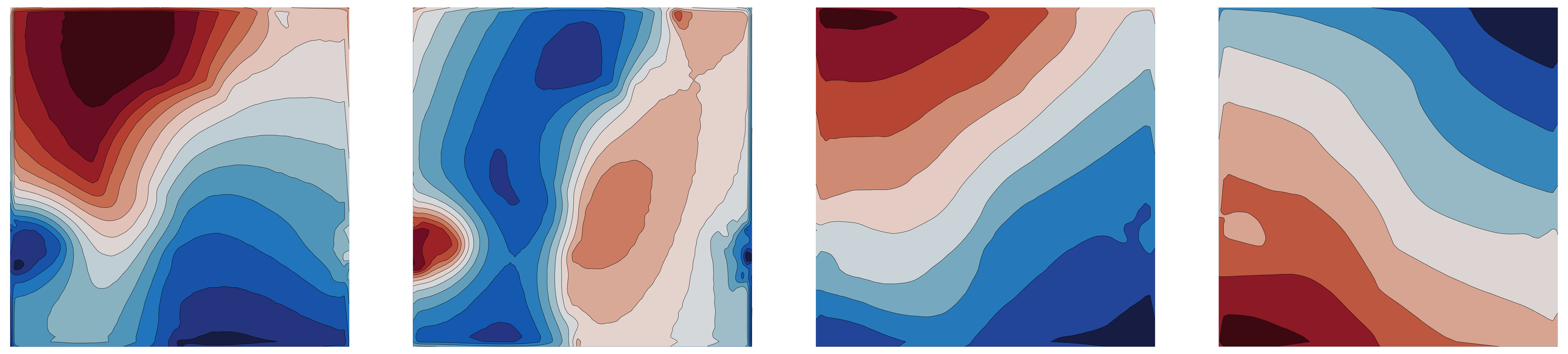
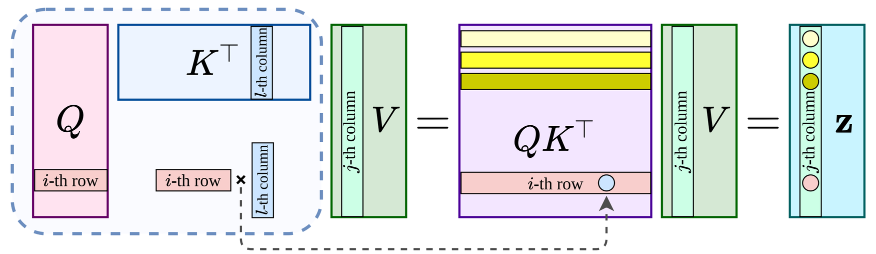


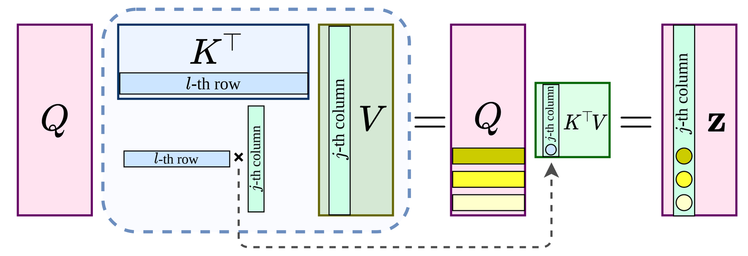
Fig. 2: Extracted latent bases during evaluation for the coefficient to solution mapping for Darcy interface problem two basis functions from the first (left two) and the fourth (right two) encoder layer in the Galerkin Transformer.

Fourier or Galerkin



$$(\boldsymbol{z}_i)_j = h\sum_{l=1}^{n} (\boldsymbol{q}_i \cdot \boldsymbol{k}_l)(\boldsymbol{v}^j)_l \approx \int_\Omega \kappa(x_i,\xi)v_j(\xi)\,\mathrm{d}\xi.$$

- Combined with the skip-connection and a diagonal scaling, this is a learnable forward propagation of the Fredholm equation of the second-kind for each basis in value. When using an explicit orthogonal expansion to seek for a better set of $\{v_j(\cdot)\}$, it is equivalent to the Nyström's method with numerical integrations [3].



$$(\boldsymbol{z}^j)_i = h\sum_{l=1}^{d} (\boldsymbol{k}^l \cdot \boldsymbol{v}^j)(\boldsymbol{q}^l)_i \approx \sum_{l=1}^{d}\left(\int_\Omega k_l(\xi)v_j(\xi)\,\mathrm{d}\xi\right) q_l(x_i).$$

- The trial space: the column space of value; the test space: the column space of key. The forward propagation can represent a learnable Petrov–Galerkin projection for every basis represented by the columns of query. Galerkin projection is the cornerstone of ALL approximation methods for operator equations in Hilbert spaces [4,5].

[3] J.-P. Berrut and M. R. Trummer, *Equivalence of Nyström's method and Fourier methods for the numerical solution of Fredholm integral equations*, Mathematics of computation 48.178 (1987), pp. 617–623

[4] P. G. Ciarlet, *Linear and nonlinear functional analysis with applications*, vol. 130, SIAM, 2013

[5] S. C. Brenner and R. Scott, *The mathematical theory of finite element methods*, vol. 15, Springer, 2008

A Ceá-type lemma for Galerkin-type attention

Theorem. *$\mathbb{Q}_h \subset \mathscr{Q}$ and $\mathbb{V}_h \subset \mathscr{V}$ are the current approximation space, suppose there exists a continuous key-to-value map that is bounded below on the discrete approximation space, i.e., the functional norm of $v \mapsto \mathfrak{b}(q,v)$ is bounded below for any q, then for g_θ consists a Galerkin attention composed with a channel reduction map*

$$\min_\theta \|f - g_\theta(\mathbf{y})\| \le \underbrace{c^{-1}}_{\|\mathfrak{b}(q,\cdot)\|_{\mathbb{V}_h'}\ge c} \underbrace{\min_{q\in\mathbb{Q}_h}\max_{v\in\mathbb{V}_h} \frac{|b(\Pi f - q, v)|}{\|v\|}}_{\textit{(Error of the Petrov-Galerkin projection)}} + \underbrace{\|f - \Pi f\|}_{\textit{(Consistency)}}.$$

- Intepretation: for an incoming "query", to deliver the best approximator in "value" (trial), the "key" space (test) has to be big enough so that for every value there is a key to unlock it. The discrete Ladyzhenskaya–Babuška–Brezzi inf-sup condition: why Transformers have capacity to generalize so well with respect to the length of the sequence!

Galerkin Transformer encoder layer

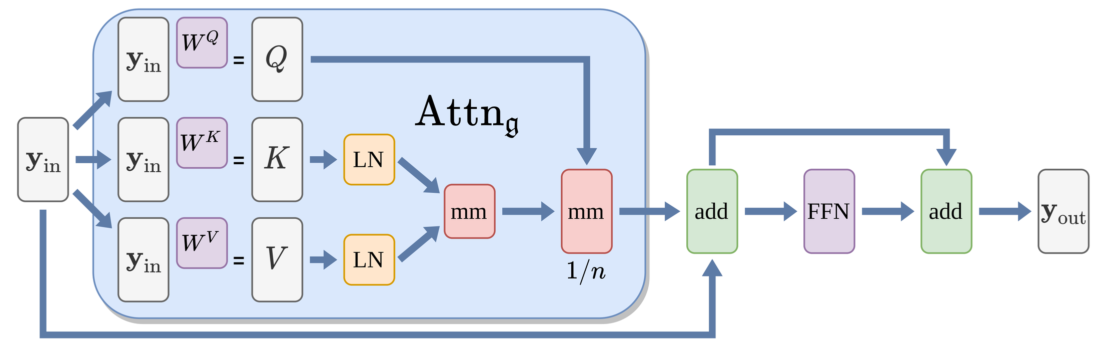


Fig. 5: Transformer encoder layer using Galerkin-type layer normalization.

- The Galerkin-type scaled dot-product attention $\text{Attn}_{\mathfrak{g}}(\mathbf{y}) := Q\big((\text{Ln}(K))^\top \text{Ln}(V)\big)/n$. The simple attention operator without softmax is

$$\text{Attn}_{\text{simple}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d}, \quad \tilde{\mathbf{y}} \leftarrow \mathbf{y} + \text{Attn}_{\mathfrak{g}}(\mathbf{y}), \quad \mathbf{y} \mapsto \tilde{\mathbf{y}} + g(\tilde{\mathbf{y}}),$$

- A Galerkin projection-like layer normalization scheme together with mesh-weighted $(1/\sqrt{n})$ instead to tame the explosive matrix product.
- Positional encoding is recurrently enriched in every encoder layer/head.
- Computational complexity is $\mathscr{O}(nd^2)$, cheaper than those with exponential feature maps (Random Feature Attention, FAVOR+ in Performer, etc.).
- A new projection weights initialization scheme inspired by the Neural ODE layer propagation scheme and the proof of the Ceá-type lemma.

Benchmark of viscous Burgers' equation

	$n = 2048$ (eval)	GFLOP/backprop	$n = 8192$ (zero-shot eval)	# params
FNO1d [2] re-implement	4.37×10^{-3}	**369.13 ± 1.81**	4.18×10^{-3}	549k
ST [1] with all tricks	2.31×10^{-3}	1876.36 ± 2.01	2.07×10^{-3}	523k
RFA [6]	1.72×10^{-2}	480.11 ± 1.74	1.91×10^{-2}	523k
FAVOR$^+$ [7]	1.58×10^{-3}	510.90 ± 25.11	1.67×10^{-3}	523k
GT with some tricks	2.45×10^{-3}	411.78 ± 1.83	2.49×10^{-3}	530k
GT with all tricks	$\mathbf{1.09\times10^{-3}}$	411.78 ± 1.83	$\mathbf{1.11\times10^{-3}}$	530k
GT 500 epochs	$\mathbf{7.79\times10^{-4}}$	411.78 ± 1.83	$\mathbf{7.90\times10^{-4}}$	530k
FNO1d 500 epochs	2.47×10^{-3}	**369.13 ± 1.81**	2.40×10^{-3}	549k
MWO [8] 500 epochs	1.86×10^{-3}	?	?	501k
XD [9] 500 epochs	9.9×10^{-3}	?	?	?

Tab. 1: Evaluation relative error/ablation study: to learn $u_0 \mapsto u(\cdot,1)$; 1024 training samples, 100 testing samples for models except for MWO and XD. All attention-based models use 4 encoder layers+2 spectral conv smoother.

[6] H. Peng, N. Pappas, D. Yogatama, R. Schwartz, N. Smith, and L. Kong, *Random Feature Attention*, in: *International Conference on Learning Representations*, 2021, URL: https://openreview.net/forum?id=QtTKTdVrFBB

[7] K. M. Choromanski et al., *Rethinking Attention with Performers*, in: *International Conference on Learning Representations (ICLR)*, 2021, URL: https://openreview.net/forum?id=Ua6zuk0WRH

[8] G. Gupta, X. Xiao, and P. Bogdan, *Multiwavelet-based Operator Learning for Differential Equations*, in: *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021, arXiv:cs.LG/2109.13459, URL: https://openreview.net/forum?id=LZDiWaC9CGL

[9] N. C. Roberts, M. Khodak, T. Dao, L. Li, C. Re, and A. Talwalkar, *Rethinking Neural Operations for Diverse Tasks*, in: *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021, arXiv:cs.LG/2103.15798, URL: https://openreview.net/forum?id=je4ymjfb5LC

References and Acknowledgments

S. Cao, *Choose a Transformer: Fourier or Galerkin* (2021), arXiv:cs.LG/2105.14995, URL: https://openreview.net/forum?id=ssohLcmn4-r, scaomath/galerkin-transformer

S. Cao is supported in part by National Science Foundation grants DMS-1913080 and DMS-2136075.

The RTX 3090 is kindly donated by Andromeda Saving Fund.

Dr. Long Chen (Univ of California Irvine) for the inspiration of and encouragement on the initial conceiving of this paper. Dr. Ari Stern (Washington Univ in St. Louis) for the help during the COVID-19 pandemic. Dr. Likai Chen (Washington Univ in St. Louis) for the invitation to the Stats and Data Sci seminar at WashU that resulted the reboot of this study. Dr. Ruchi Guo (Univ of California Irvine) and Dr. Yuanzhe Xi (Emory Univ) for the invaluable feedbacks on the choice of the numerical experiments. Zongyi Li (Caltech) for sharing some early dev code in the updated PyTorch `fft` interface and the comments on the viscosity of the Burgers' equation.