REPORT ON THE PAPER "NONLOCALITY AND NONLINEARITY IMPLIES UNIVERSALITY IN OPERATOR LEARNING"

DEVANSH TRIPATHI
ETH Zürich

Abstract. Neural operator architecture approximate operators between infinite dimensional Banach spaces of function. The paper discuss the basic questions about the requirements for universal approximation of neural operator and provide conditions under which neural operators are universal approximators. Author argued that the general approximation of operators between spaces of functions must be both *nonlocal* and *nonlinear*.

A popular variant of neural operators is the Fourier neural operator (FNO). Proving universal approximation theorem for FNOs is based on using unbounded number of Fourier modes, this work challenges this point of view and provide a novel minimal architecture called "averaging neural operator" (ANO) and its analysis showed that ANO is also a universal approximator. Only spatial average is taken as nonlocal ingredient which corresponds to retaning only a single Fourier mode in the case of FNO contrasts to unbounded number of modes.

1. Introduction

The task that we are trying to achieve with the help of neural networks is to approximate the underlying operator, which defines a mapping between two infinite-dimensional Banach spaces of functions. Neural operator generalizs the underlying framework of neural network to infinite-dimensional setting and learn such operators from the data. The wide range of neural operators introduced in [1, 4] are defined in analogy with neural networks, but appends weight matrices in the hidden layers with additional linear integral operators acting on the input functions. Special cases of this framework includes Fourier neural network (FNO) [6] in which the reliance on a Fourier basis limits the basic form of the FNO to periodic geometries, although, in that setting, use of fast Fourier transform (FFT) allows for efficient computations with total number of Fourier components limited by the grid resolution. Extension of FNO, called Neural Operator on Riemannian Manifold (NORM) [2], generalizes FNO to use arbitrary orthogonal eigenfunctions of the Laplace-Beltrami operator on any given spatial domain. The approach mentioned in the paper is closely related to low rank neural operator [4] which, however, has a more complicated architecture that proposed in this paper.

Seminar für Angewandte Mathematik, HG E 62.2, Rämistrasse 101, 8092 Zürich, Switzerland devansh.tripathi@sam.math.ethz.ch.

In the paper author argued that universal approximation can be obtained in general geometries, with nonlocality along with nonlinearity. Nonlocality has been introduced using only a low-rank operator of fixed finite rank, and is not restricted to periodic domains.

Fourier neural operators are already nonlocal and they introduce it via the addition of a nonlocal operator in each hidden layer layer, which acts on the Fourier modes of the input function by matrix multiplication. FNOs are generally implemented with a first layer which lifts the input, a scalar or vector-valued function, to a vector-valued function where the vector dimension (also called model width) is much higher than that of input function itself. It has been showed in literature that increasing the number of channels (model width) rather than to retain more Fourier modes in the architecture is more beneficial in certain circumstances [5].

The author has proposed a underlying framework to many neural operators called averaging neural operator (ANO). The ANO is build upon two minimal ingredients, nonlinearity by composition of shallow neural networks, and nonlocality via a spatial average. Author has deduced many universal approximation theorems for the ANO.

1.1. **Neural Operator.** Let $\Omega \subset \mathbb{R}^d$ denote a bounded domain (or potentially a manifold) and let $\mathcal{X}(\Omega;\mathbb{R}^o)$,$\mathcal{Y}(\Omega;\mathbb{R}^o)$ and $\mathcal{V}(\Omega;\mathbb{R}^o)$ denote Banach spaces of $\mathbb{R}^o-$ valued functions over Ω. The *nonlocal neural operator* (NNO) is defined as a mapping

$$\Psi\colon \mathcal{X}(\Omega;\mathbb{R}^k) \to \mathcal{Y}(\Omega;\mathbb{R}^k)$$

which can be written as composition of the form $\Psi = \mathcal{Q} \circ \mathcal{L}_L \circ \dots \mathcal{L}_1 \circ \mathcal{R}$ where $\mathcal{R}$ is lifting layer, $\mathcal{L}_l, l = 1, \dots L$ are hidden layers and $\mathcal{Q}$ is projection layer. Given a channel dimension d_c, the **lifting layer** $\mathcal{R}$ and **projection layer** $\mathcal{Q}$ are given by a mapping respectively:

$$\mathcal{R}\colon \mathcal{X}(\Omega;\mathbb{R}^k) \to \mathcal{V}(\Omega;\mathbb{R}^{d_c}),\ u(x) \mapsto R(u(x), x), \tag{1}$$

$$\mathcal{Q}: \mathcal{V}(\Omega;\mathbb{R}^{d_c}) \to \mathcal{Y}(\Omega;\mathbb{R}^{d_c}),\ v(x) \mapsto Q(v(x), x) \tag{2}$$

where $R\colon \mathbb{R}^k \times \Omega \to \mathbb{R}^{d_c}$ and $Q : \mathbb{R}^{d_c} \times \Omega \to \mathbb{R}^{k'}$ are learnable neural network acting between finite dimensional Euclidean spaces. For $l = 1, \dots, L$ (the number of **hidden layers**) and for $m = 0, \dots, M$ (the number of modes) choose functions $\psi_{l,m}, \phi_{l,m}\colon \Omega \to \mathbb{R}^{d_c}$. For $l = 1, \dots, L$, each hidden layer $\mathcal{L}_l$ is the mapping $\mathcal{V}(\Omega;\mathbb{R}^{d_c}) \to \mathcal{V}(\Omega;\mathbb{R}^{d_c})$ of the form:

$$(\mathcal{L}_l v)(x) := \sigma\left(W_l v(x) + b_l + \sum_{m=0}^{M} \langle T_{l,m} v, \psi_{l,m}\rangle_{L^2(\Omega;\mathbb{R}^{d_c})} \phi_{l,m}(x)\right)$$

where

(1) $W_l, T_{l,m} \in \mathbb{R}^{d_c \times d_c}$ and bias $b_l \in \mathbb{R}^{d_c}$ are the learnable parameters/matices.
(2) $\langle T_{l,m} v, \psi_{l,m}\rangle_{L^2(\Omega;\mathbb{R}^{d_c})}$ is a inner product in L^2- function space. It measure the contribution of $\psi_{l,m}$ in $T_{l,m}$.

Note.

(1) The L^2 space is space of square integrable functions such that $L^2(\Omega;\mathbb{R}^{d_c}) := \left\{f \mid \int_\Omega \|f(x)\|^2 dx < \infty\right\}$.
(2) The form of lifting and projection layer allows for *positional encoding*.

(3) This general framework reduces to the FNO in a periodic geometry and if the expansion function are choosen as Fourier basis functions, indexed by m and independent of l.

The central questions that author posses: "which minimal assumptions have to be imposed on the expansion functions $\psi_{l,m}$ and $\phi_{l,m}$, to ensure universal approximation of the resulting architecture?"

Assumptions on activation function: $\sigma : \mathbb{R} \to \mathbb{R}$ is assumed to be smooth, $\sigma \in C^\infty(\mathbb{R})$, nonpolynomial and Lipschitz continuous. It acts as a Nemitskii-operator, component-wise on inputs.

2. Averaging suffices for Universal Approximation

Author proposes a new architecture called averaging neural operator (ANO) and shows that only a simple averaging as nonlocality suffices for universality. The ANO is the subclass of many instantiations of general NNO architecture, and hence implies universality for general NNOs as corollaries. When the domain is periodic, the resulting ANO becomes a special case of FNO when only zeroth Fourier mode is retained.

2.1. Nonlinearity and Nonlocality.

In case of ordinary neural networks, nonlinearity alone suffies for universality but in case of neural operators, it does not alone suffies. To support this, we have a following example: a neural network with a single layer give rise to nonlinear operator, which maps an input function to an output function by composition,

$$u(x) \mapsto \sigma(Wu(x) + b).$$

Despite being nonlinear, it can be shown that they are not universal. For example, such mappings are not able to approximate even simple operators $\Psi^\dagger$ with a *nonlocal* dependence on the input, such as the shift operator $\Psi^\dagger(u)(x) := u(x + h)$ for fixed $h \neq 0$. This shows the need of nonlocality.

2.2. Averaging Neural Operator: a Special Subclass of the NNO.

The author has defined a special subclass of the NNO, which combines nonlinearity by composition with nonlocality by averaging. We define a special subclass of hidden layers of the form:

$$\mathcal{L} : \mathcal{V}(\Omega; \mathbb{R}^{d_c}) \to \mathcal{V}(\Omega; \mathbb{R}^{d_c}),\ \mathcal{L}(v)(x) := \sigma\left(Wv(x) + b + \fint_\Omega v(y)dy\right) \tag{3}$$

The author has taken up the case of *single hidden layer* hence the following architecture:

$$\Psi : \mathcal{V}(\Omega; \mathbb{R}^{d_c}) \to \mathcal{V}(\Omega; \mathbb{R}^{d_c}),\ \Psi(u) = \mathcal{Q} \circ \mathcal{L} \circ \mathcal{R}(u),$$

where lifting and projection maps $\mathcal{R}$ and $\mathcal{Q}$ are given as in above equations with R and Q be single-hidden layer neural networks respectively of width d_c.

Parameters of ANO. Due to its minimal structure, the ANO depends on only one *hyperparameter;* the lifting dimension d_c. The *tunable parameters* of ANO are represented by the weight matrix $W \in \mathbb{R}^{d_c \times d_c}$ and bias $b \in \mathbb{R}^{d_c}$ in the hidden layer $\mathcal{L}$, and the internal weights and biases of the ordinary neural network R and Q defined in lifting and projection layers, respectively.

2.3. **Universal Approximation.**

Theorem 1. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. For given integers $s, s' \geq 0$ let $\Psi : C^s(\overline{\Omega}; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ be continuous operator, and fix a compact set $K \subset C^s(\overline{\Omega}; \mathbb{R}^k)$. Then for any $\epsilon > 0$, there exists an averaging neural operator $\Psi : K \subset C^s(\overline{\Omega}; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ such that*

$$\sup_{u \in K} \|\Psi^\dagger(u) - \Psi(u)\|_{C^{s'}} \leq \epsilon.$$

The above result is in the space of continuously differentiable functions. Author also has a corresponding result in the scale of Sobolov spaces $W^{s,p}$:

Theorem 2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. For given integers $s, s' \geq 0$, and reals $p, p' \in [1, \infty)$, let $\Psi^\dagger : W^{s,p}(\Omega; \mathbb{R}^k) \to W^{s',p'}(\Omega; \mathbb{R}^{k'})$ be a continuous operator. Fix a compact set $K \subset W^{s,p}(\Omega; \mathbb{R}^k)$ of bounded functions, $\sup_{u \in K} \|u\|_{L^\infty} < \infty$. Then for any $\epsilon > 0$, there exists an averaging neural operator $\Psi : W^{s,p}(\Omega; \mathbb{R}^k) \to W^{s',p'}(\Omega; \mathbb{R}^{k'})$ such that*

$$\sup_{u \in K} \|\Psi^\dagger(u) - \Psi(u)\|_{W^{s',p'}} \leq \epsilon.$$

The consequences of the results is that any more general NNP architecture hidden layers of the form (3), is universal as long as an average can be represented; then the resulting NNO reduces to ANO with a specific setting of tunable weights (by setting certain parameters to zero).

Remark 1. Above two theorem show that ANO is universal i approximating a large class of operators, uniformly over a compact set of input functions K. But in practice, neural operators are trained by minimizing an empirical loss:

$$\mathcal{L}(\Psi) = \frac{1}{N} \sum_{n=1}^{N} \|\Psi^\dagger(u_n) - \Psi(u_n)\|^2_{W^{s',p'}(\Omega)}, \; u_1, \ldots u_N \sim \mu,$$

where data are iid random samples form an underlying input probability measure μ. A popular choice is sampling input function from a Gaussian random field. In this case, the set of input functions is no longer bounded, and hence not compact.

In this case when μ does not have a compact support, we can derive these results with a cut-off argument. For given $\epsilon > 0$, it is possible to show that the existence of Ψ, such that

$$\mathbb{E}_{u \sim \mu}[\|\Psi(u) - \Psi^\dagger(u)\|^2_{W^{s',p'}}] < \epsilon.$$

2.4. **Intuition.**

Encoder-Decoder Structure. The ANO architecture has hidden encoder-decoder structure. This structure can be obtained when setting the matrix W and bias b in the hidden layer (3) zero, in which case we note that the mapping $\mathcal{L} \circ \mathcal{R} : \mathcal{X} \to \mathbb{R}^{d_c}, u \mapsto \mathcal{L} \circ \mathcal{R}(u)$ can be though of as a nonlinear encoding of the input function and encode it to a (constant) vector $v \in \mathbb{R}^{d_c}$ while the projection mapping that takes $v \mapsto Q(v, .)$ is a nonlinear decoding of the corresponding output function $\Psi^\dagger(u) = Q(u, x)$.

The Role of Positional Encoding. Author has argued that some positional encoding ("explicit x-dependence") is necessary for universality. Positional encoding helps in breaking translational equivariance i.e. FNO Ψ does not commute with the shift operator. Breaking translational equivariance is necessary because the PDEs with non-constant coefficients have solution operator which are non-translational equivariant. There is one more way to break translational equivariance that is to have bias function dependent on $x, b = b(x)$. It was used in previous work by the author [3] but in practice positional encoding is explicitly added in the input layer $\mathcal{R}$.

Explicit x-dependence in the output layer is not necessary since there are pointwise matrix multiplications involved in the hidden layers and they provide a mechanism (same as "skip connections") to forward positional encoding information to the output layer.

Remark 2 (Skip Connections)**.** Skipping connections basically means bypass one or more layer of the neural network. It adds the output of layer(l) to the output of layer($l + n$):

$$y = F(x) + x$$

where:

- x is the input.
- $F(x)$ is the transformation applied to skip layer(s).
- y is the output after the skip.

2.5. **Sketch of the Proof of Universality.** For the sake of completeness, theorem statement has been stated again below:

Theorem 3. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. For given integers $s, s' \geq 0$ let $\Psi : C^s(\overline{\Omega}; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ be continuous operator, and fix a compact set $K \subset C^s(\overline{\Omega}; \mathbb{R}^k)$. Then for any $\epsilon > 0$, there exists an averaging neural operator $\Psi : K \subset C^s(\overline{\Omega}; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ such that*

$$\sup_{u \in K} \|\Psi^\dagger(u) - \Psi(u)\|_{C^{s'}} \leq \epsilon.$$

Explanation and definition of terms:

Lipschitz Domain. $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain and $\overline{\Omega}$ is compact (Heine-Borel theorem). The boundary $\delta\Omega = \overline{\Omega} \backslash \Omega$ is at least "Lipschitz regular"(locally it is the graph of a Lipschitz function). In particular, any bounded domain with piecewise smooth boundary is a Lipschitz domain.

Function Extension for Lipschitz Domains. We will be using this extension property in the proof. Given a function $u : \Omega \to \mathbb{R}$, defined on a domain $\Omega \subset \mathbb{R}^d$, to a function $u : \mathbb{R}^d \to \mathbb{R}$, defined on all of $\mathbb{R}^d$. The following lemma shows that this is possible, while preserving smoothness of u.

Lemma 1 (Periodic extension operator)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. There exists a continuous, linear operator $\mathcal{E} : W^{s,p}(\Omega) \to W^{(}_{per}s, p)(B)$ for any $s \geq 0$ and $p \in [1, \infty]$, where $B \subset \mathbb{R}^d$ is a bounded hypercube containing $\Omega \subset B$, such that for any $u \in W^{s,p}(\Omega)$:*

References

[1] A. Anandkumar, K. Azizzadenesheli, et al. Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[2] Guannan Chen, Xuan Liu, Qi Meng, Long Chen, Chunmei Liu, and Yulan Li. Learning neural operators on riemannian manifolds. 2023. Preprint.

[3] Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021.

[4] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

[5] Samuel Lanthaler, Roberto Molinaro, Philipp Hadorn, and Siddhartha Mishra. Nonlinear reconstruction for operator learning of pdes with discontinuities. *arXiv preprint arXiv:2210.01074*, 2022. URL `https://arxiv.org/abs/2210.01074`. Preprint.

[6] Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.